Human Evaluation of NLP System Quality

INLG Tutorial, 24th September 2024

Unit 5: Statistical Analysis of Results

Link to Unit 5 Resources

Overview

Unit 5: Analysis of Results

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Introduction to statistics – hypothesis testing
3. Connection to pre-registration (statistical power analysis)
4. Multiple hypothesis testing
5. Annotator reliability and representativeness
6. Post-hoc tests
7. Practical coding session
8. Unit summary
9. References

Prerequisites and connections with other units

- Prerequisite(s) of Unit 5: Units 2, 3, and 4
- Topics: research questions, hypothesis testing, and evaluation modes introduced
- Unit 5 is a prerequisite of Units 6–8, as it introduces design choices that later units will assume knowledge of probability

Unit aims and learning outcomes

- The aims of Unit 5 are:
  - To revisit the analysis of results from a statistical perspective
  - To present null hypothesis significance testing, statistical significance tests, and power analysis
  - Also covered are data transformations, issues with data, and perspective solutions
- After completion of the unit, participants will be able to:
  - Understand hypothesis testing and formulation
  - Understand and make choices about which tests to use
  - Gain an understanding of pre-registration and confirmatory vs exploratory hypothesis testing

Overview

Unit 5: Analysis of Results

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Introduction to statistics – hypothesis testing
3. Connection to pre-registration (statistical power analysis)
4. Multiple hypothesis testing
5. Annotator reliability and representativeness
6. Post-hoc tests
7. Practical coding session
8. Unit summary
9. References

Preliminary data analysis

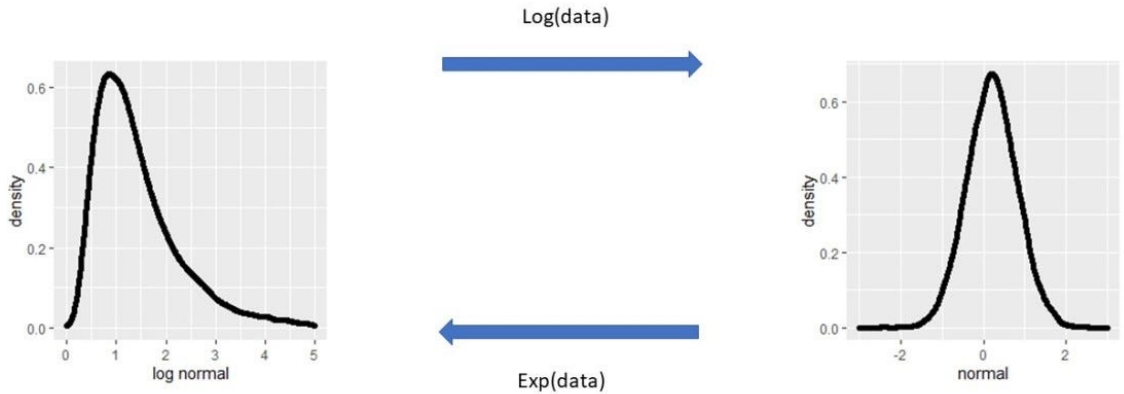
- Histogram and co-variance plots (scatterplot)
- Error detection: Outlier detection, missing data, deduplication
- Annotation agreement
- Normality tests
- Data transformations (e.g., binning, scaling, median aggregation)

If this an exploratory study

- Exploratory Data Analysis (EDA)
- Model free assessment
- Hypothesis creation

Preliminary data analysis

- Histogram and co-variance plots (scatterplot)
- Error detection: Outlier detection, missing data, deduplication
- Annotation agreement
- Normality tests
- Data transformations (e.g., binning, scaling, median aggregation)



Preliminary data analysis – Look at Data

```
1 ratings_df.head()
```

id	mr	team	text	category	type	bleu	meteor	ter	systemtype	triplesize	fluency	grammar	semantics
1	(29075)_1950_DA | discoverer | Carl_A._Wirtanen	adapt	the 29075 club is the dictcoverer, carl a. wir...	CelestialBody	unseen	0.041	0.185956	90.909	neural	1triple	1.666667	1.666667	1.333333
2	(29075)_1950_DA | discoverer | Carl_A._Wirtanen	baseline	the administrative government is governed by t...	CelestialBody	unseen	0.034	0.046764	90.909	neural	1triple	2.750000	2.750000	1.000000
3	(29075)_1950_DA | discoverer | Carl_A._Wirtanen	melbourne	1950 da is carl a. wirtanen.	CelestialBody	unseen	0.066	0.320360	81.818	neural	1triple	2.000000	2.333333	1.000000

Source: WebNLG 2017 human evaluations

Preliminary data analysis – Look at Data

```
1 ratings_df.head()
```

id	mr	team	text	category	type	bleu	meteor	ter	systemtype	triplesize	fluency	grammar	semantics
1	(29075)_1950_DA | discoverer | Carl_A._Wirtanen	adapt	the 29075 club is the dictcoverer, carl a. wir...	CelestialBody	unseen	0.041	0.185956	90.909	neural	1triple	1.666667	1.666667	1.333333

```
1 ratings_df.tail()
```

id	mr	team	text	category	type	bleu	meteor	ter	systemtype	triplesize	fluency	grammar	semantics
2226	William_Anders | status | "Retired" William...	tilburg-pipe	william anders graduated from afit in 1962 him...	Astronaut	seen	0.131	0.312793	59.091	template	2triple	1.666667	1.666667	2.333333
2227	William_Anders | status | "Retired" William...	tilburg-smt	william anders is now retired. william anders ...	Astronaut	seen	0.085	0.434852	77.273	smt	2triple	2.333333	2.666667	3.000000
2228	William_Anders | status | "Retired" William...	upf-forge	william anders, who graduated from afit in 196...	Astronaut	seen	0.124	0.412168	68.182	template	2triple	3.000000	3.000000	3.000000

Preliminary data analysis – Look at Data

```
1 ratings_df.head()
```

id	mr	team	text	category	type	bleu	meteor	ter	systemtype	triplesize	fluency	grammar	semantics
1	(29075)_1950_DA | discoverer | Carl_A._Wirtanen	adapt	the 29075 club is the dictcoverer, carl a. wir...	CelestialBody	unseen	0.041	0.185956	90.909	neural	1triple	1.666667	1.666667	1.333333

```
1 ratings_df.tail()
```

id	mr	team	text	category	type	bleu	meteor	ter	systemtype	triplesize	fluency	grammar	semantics

```
1 ratings_df.sample(10)
```

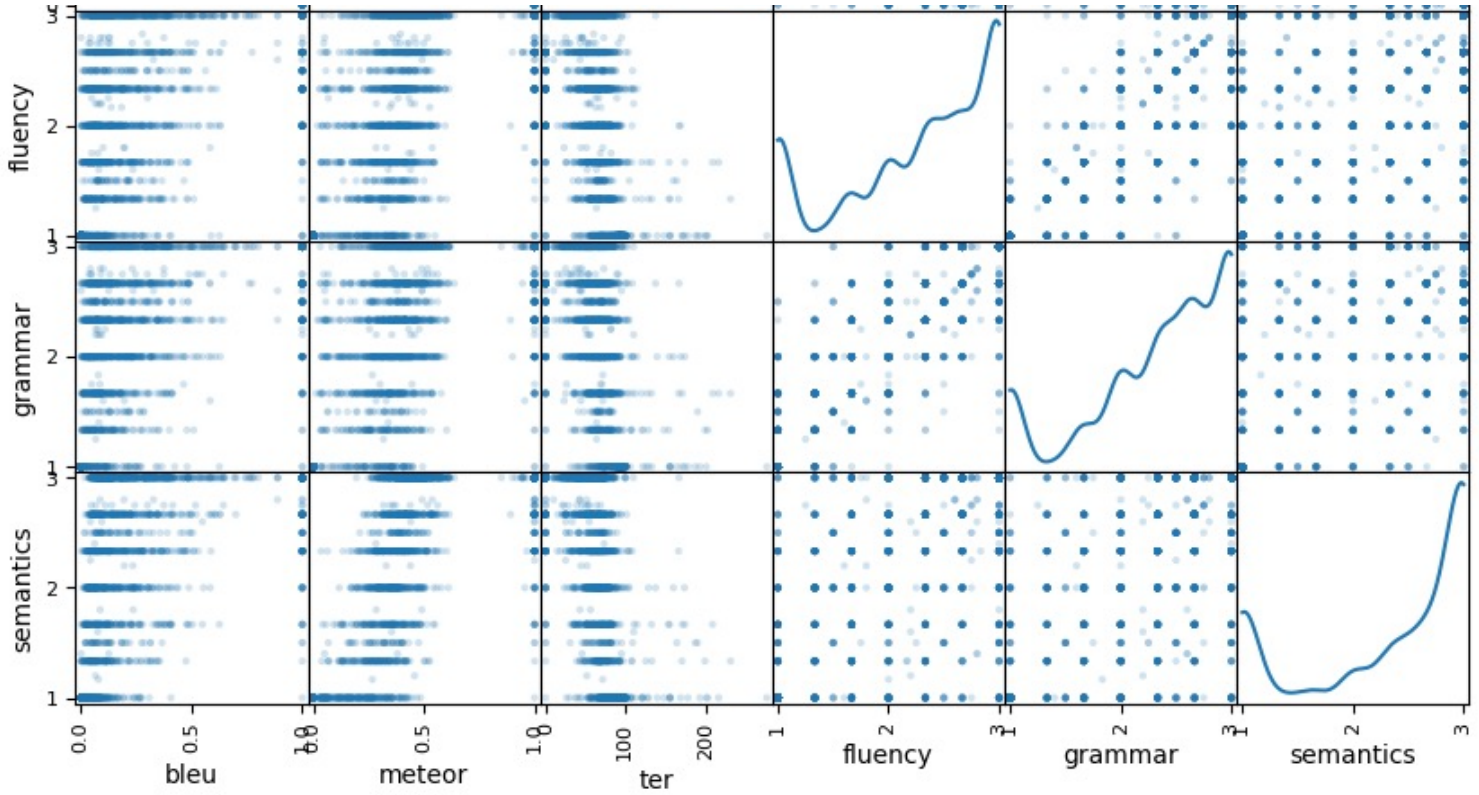
id	mr	team	text	category	type	bleu	meteor	ter	systemtype	triplesize	fluency	grammar	semantic
1414	Bandeja_paisa | mainIngredients | "red beans, ...	pkuwriter	bandeja paisa is a dish commonly found in colo...	Food	seen	0.070	0.610397	61.047	neural	4triple	2.666667	2.666667	2.00000
1581	Buzz_Aldrin | birthPlace | Glen_Ridge,_New_Jer...	adapt	buzz aldrin was born in glen ridge, essex coun...	Astronaut	seen	0.288	0.508471	51.613	neural	2triple	3.000000	3.000000	3.00000
1008	Allama_Iqbal_International_Airport | location ...	upf-forge	allama iqbal international airport, which	Airport	seen	0.137	0.370921	65.323	template	5triple	2.000000	2.666667	3.00000

Preliminary data analysis – Look at Data

```
1 ratings_df.describe()
```

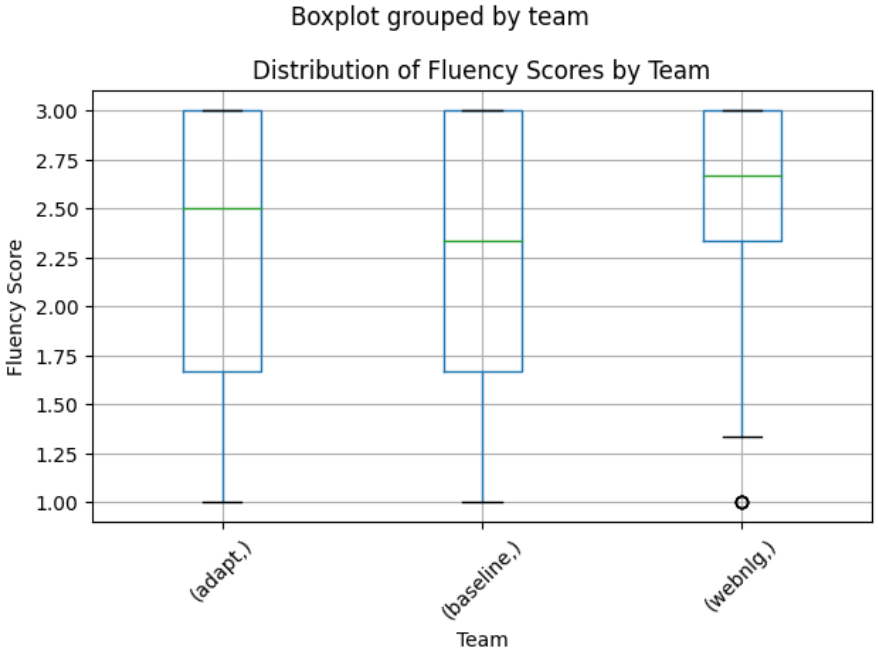
	bleu	meteor	ter	fluency	grammar	semantics
count	2230.000000	2230.000000	2230.000000	2230.000000	2230.000000	2230.000000
mean	0.241391	0.418217	59.453191	2.131345	2.223378	2.199709
std	0.289667	0.272278	30.181619	0.704540	0.675336	0.756654
min	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
25%	0.059000	0.281395	45.455000	1.666667	1.666667	1.500000
50%	0.129000	0.385286	62.791000	2.333333	2.333333	2.333333
75%	0.294000	0.485533	76.250750	2.666667	2.666667	3.000000
max	1.000000	1.000000	276.596000	3.000000	3.000000	3.000000

Preliminary data analysis – Visualize Data



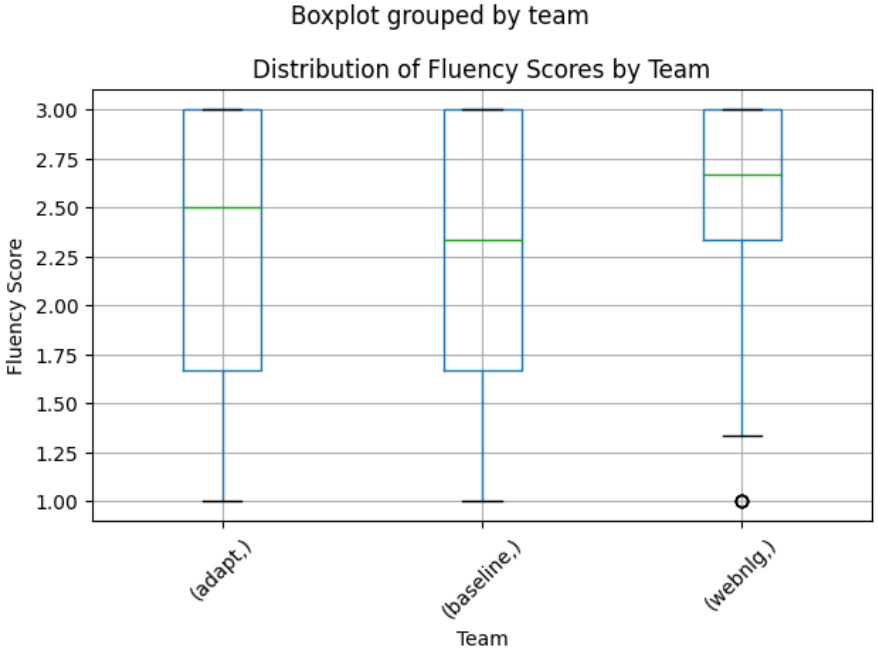
Model free evidence

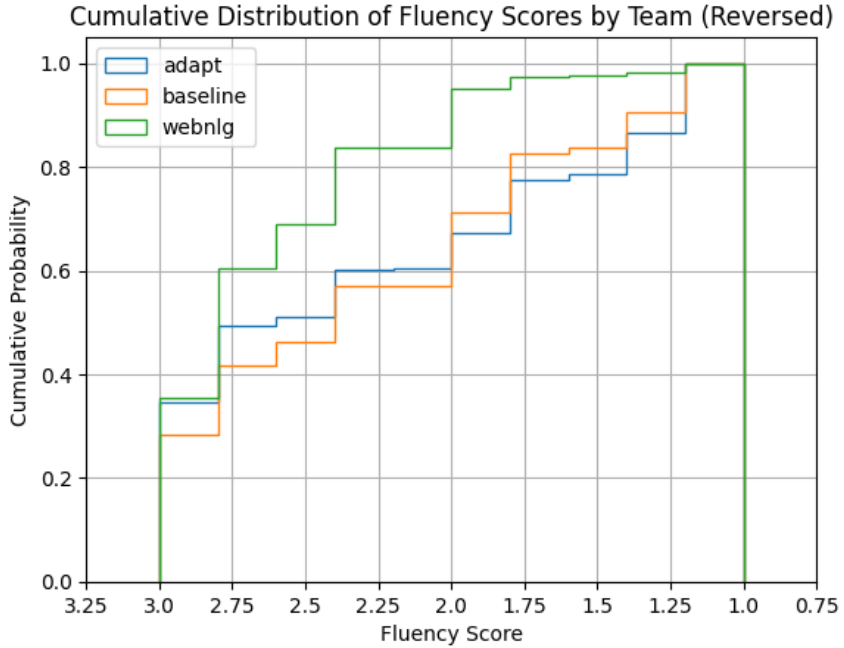
Model-free evidence - trends and patterns observed directly from the data, without fitting them into a specific theoretical model



Model free evidence

Model-free evidence - trends and patterns observed directly from the data, without fitting them into a specific theoretical model





Answering the research question

- The standard approach to answering such questions is **null hypothesis testing** in a form mainly due to Fisher (1935).
- Our **null hypothesis** is that there is no difference in terms of E_m between systems in our set.
- We look for strong enough evidence to allow us to conclude, on the basis of our test data set *s*, samples of system outputs o_i^s and values of E_m computed on them, that we are wrong in our assumption that the null hypothesis is correct, and to conclude instead that there is in fact a difference.
- We test the strength of the evidence via a **statistical hypothesis test** involving the calculation of a **test statistic**.
- Most commonly, a **p-value** computed from the test statistic is taken to indicate the level of significance.
- The p-value expresses the probability under the null hypothesis of obtaining a test statistic value that is at least as extreme as the one actually observed.
- Hence a small p-value means there is little overall chance of obtaining a result like the one observed if the null hypothesis (there is no difference between the systems in terms of E_m) is true.

Hypothesis Testing

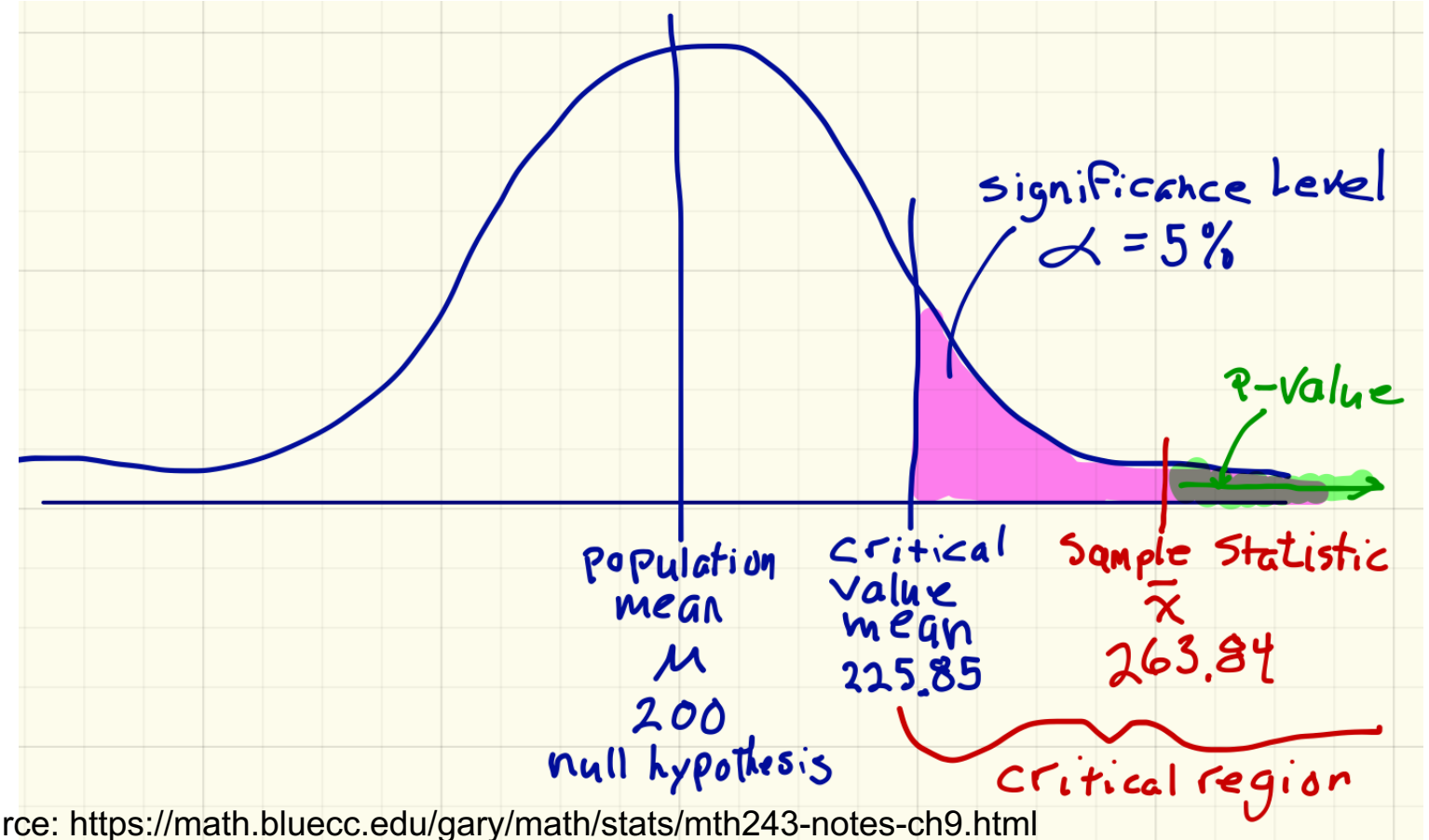
First, we'll need a bunch of definitions:

- Hypothesis - an assumption about the distribution of a random variable in a population
- Maintained hypothesis - one that cannot or will not be tested
- Testable hypothesis - one that can be tested using evidence from a random sample
- Null hypothesis, H_o - the one that will be tested
- Alternative hypothesis, H_a - a possibility (or series of possibilities) other than the null

Hypothesis Testing

- Our **null hypothesis (H_o)** is that there is no difference in terms of E_m between systems in our set
- **Alternative hypothesis (H_a)** a system outperforms other system(s)
- **Significance Level (α)**: This is the threshold for deciding whether to reject the null hypothesis
- **Test Statistic**: A numerical value calculated from the sample data, which is used to determine whether to reject the null hypothesis. The type of test statistic depends on the nature of the data and the hypothesis being tested
- **P-Value**: The probability of observing the test statistic or something more extreme, given that the null hypothesis is true

Hypothesis Testing



Source: https://math.bluecc.edu/gary/math/stats/mth243-notes-ch9.html

Hypothesis Testing

- Our **null hypothesis (H_o)** is that there is no difference in terms of E_m between systems in our set
- **Alternative hypothesis (H_a)** a system outperforms other system(s)
- **Significance Level (α)**: This is the threshold for deciding whether to reject the null hypothesis
- **Test Statistic**: A numerical value calculated from the sample data, which is used to determine whether to reject the null hypothesis. The type of test statistic depends on the nature of the data and the hypothesis being tested
- **P-Value**: The probability of observing the test statistic or something more extreme, given that the null hypothesis is true

Hypothesis Testing

Steps in Hypothesis Testing:

1. Formulate Hypotheses: Define the null and alternative hypotheses.
2. Decide on the test statistic (use prior assumptions or pilot data).
3. Choose Significance Level (α): Determine the level of risk you're willing to take. (0.05 means 1 out of 20 times reject the hull hypothesis)
4. Collect Data: Gather a sample and calculate the appropriate test statistic.
5. Calculate the Test Statistic: Use the sample data to calculate the statistic.
6. Determine the P-Value: Find the probability associated with the test statistic.
7. Make a Decision: Compare the p-value to α and decide whether to reject or not reject the null hypothesis.
8. Draw a Conclusion: State the result in the context of the original problem.

Hypothesis Testing

- Formulating the null hypothesis and alternative hypothesis
- There can be multiple types of alternatives:
  - System ordering is strict e.g., A>B>C
  - System ordering allows for ties A>B>=C
  - Simple superiority A> B and A>C
- Multiple Hypothesis tests: e.g., outperform over multiple datasets or across multiple dimensions

Hypothesis Testing

Mistakes can be made …

- We can “reject” a null that is true or “accept” a null that is false
- We want to set up our hypothesis test to analyze and control these errors
- We first need a taxonomy:

	H_0 true	H_0 false
accept H_0	No error	Type II error
reject H_0	Type I error	No error

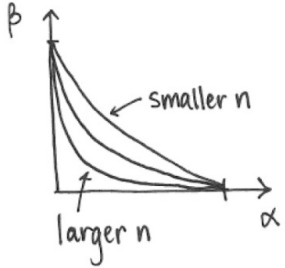
Source: https://www.youtube.com/watch?v=uHUOwZmfUzY

Hypothesis Testing

Null Hypothesis Significance Testing

- Type I error – rejecting the null hypothesis when it is true
- Type II error – not rejecting the null hypothesis when the alternative is true
- Significance level (α) – probability of making type I error
- Significance Power (β)– probability of not making type II error

What happens as n increases or decreases?



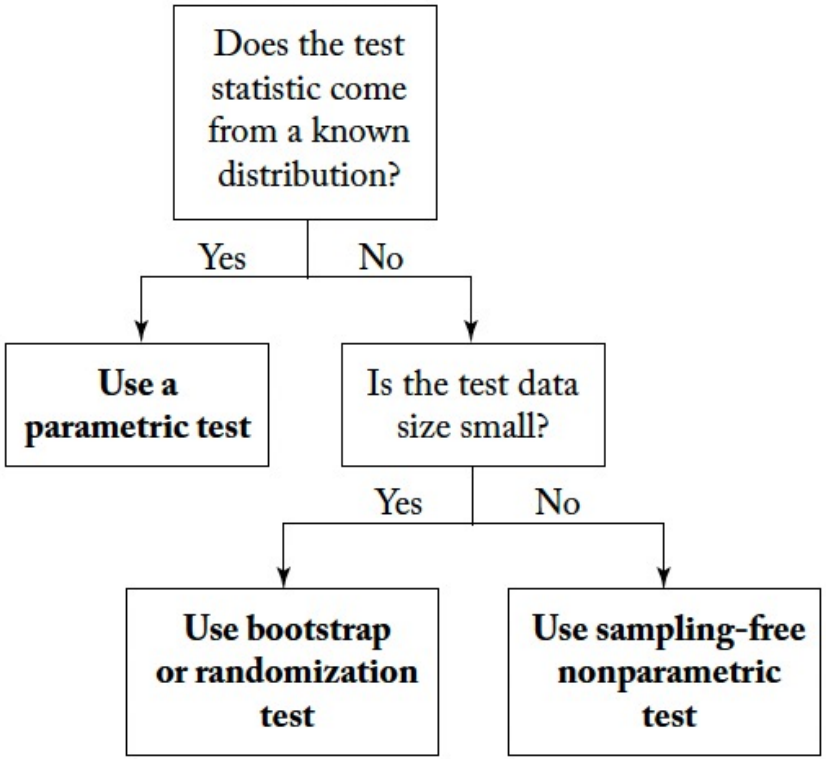
There are other error types

Type D – sign error

Type M – magnitude error

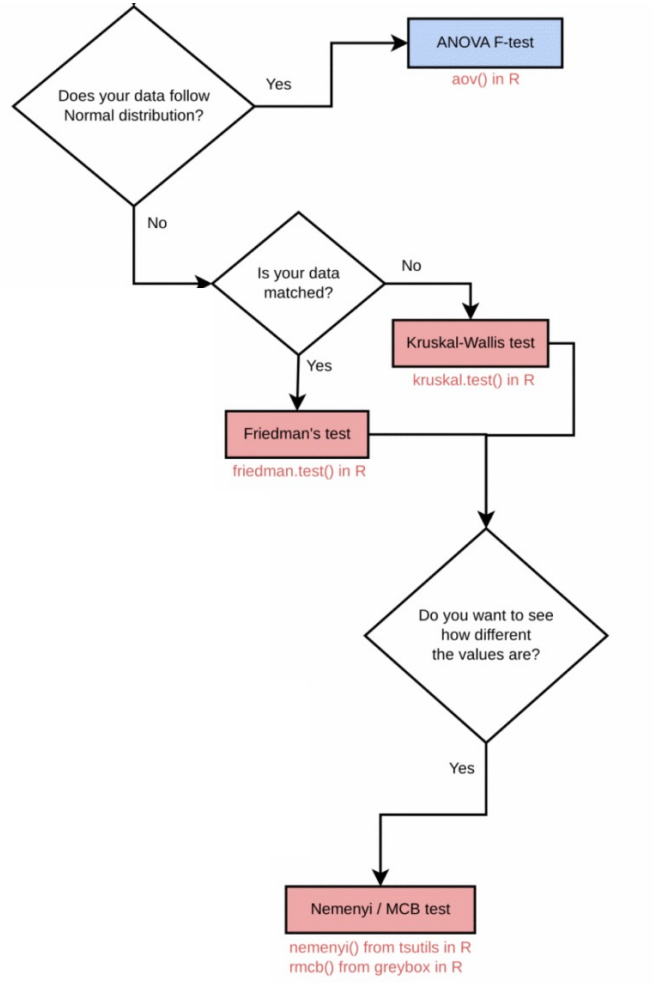
Source: https://www.youtube.com/watch?v=uHUOwZmfUzY

Picking the right statistic

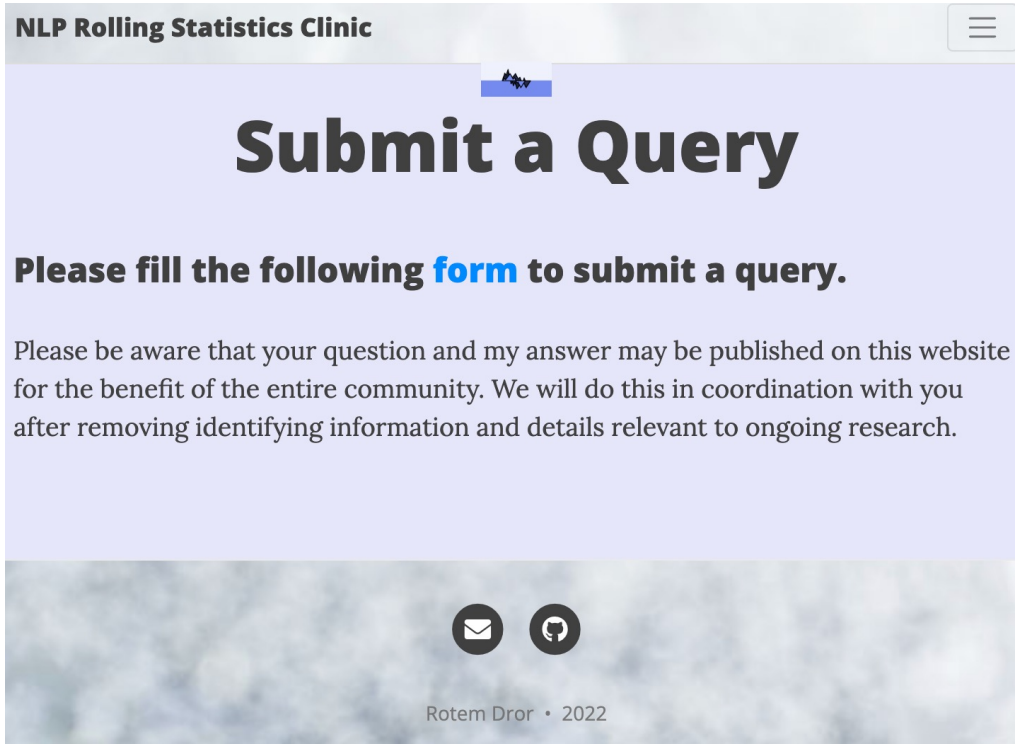


Source: The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing

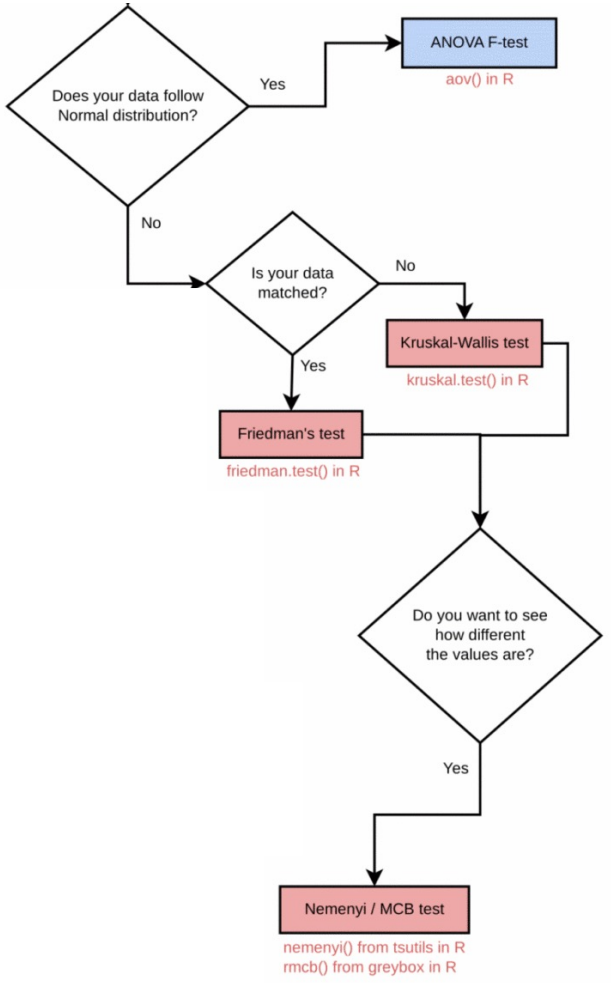
Picking the right statistic



Picking the right statistic



https://nlpstatclinic.github.io/query/



Test Statistic Assumptions (e.g., Kruskal-Wallis)

- Test statistics come with assumptions
- Kruskal – Wallis (rank based test)
  - **Assumption #1:** Your **dependent variable** should be measured at the **ordinal** or **continuous level** (i.e., **interval** or **ratio**)
  - **Assumption #2:** Your **independent variable** should consist of **two or more categorical**, **independent groups**
  - **Assumption #3:** You should have **independence of observations**, which means that there is no relationship between the observations in each group or between the groups themselves
  - **Assumption #4:** The **distributions** in each group (i.e., the distribution of scores for each group of the independent variable) have the **same shape** (which also means the **same variability**)

Overview

Unit 5: Analysis of Results

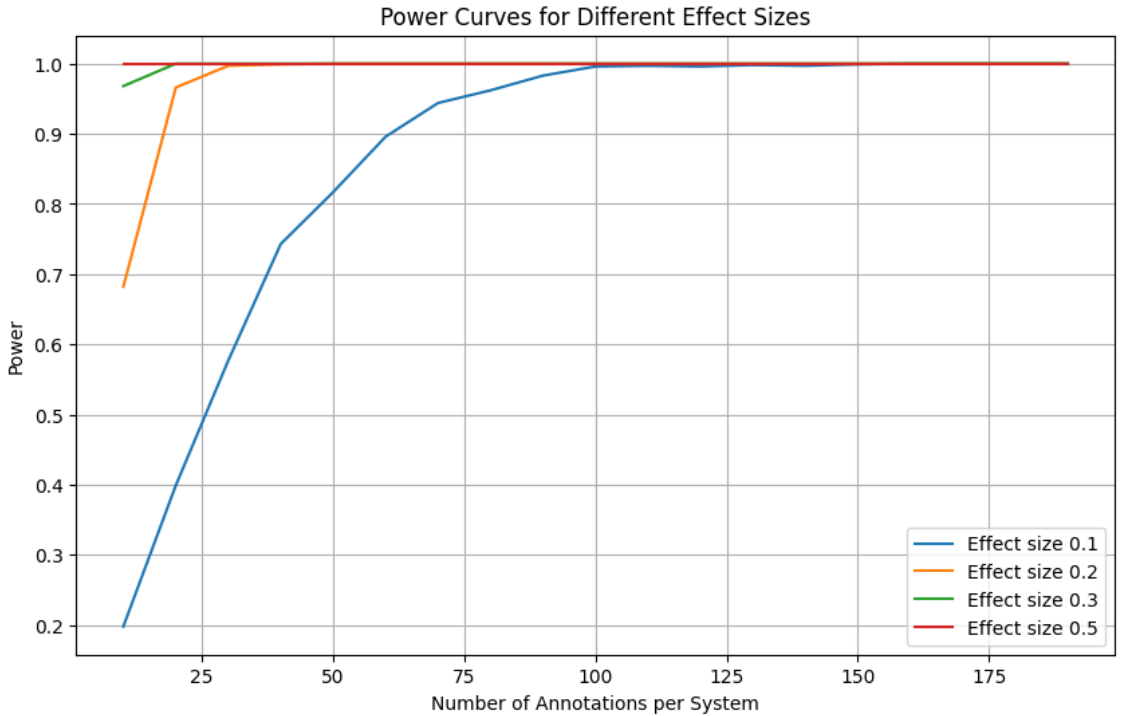
1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Introduction to statistics – hypothesis testing
3. Connection to pre-registration (statistical power analysis)
4. Multiple hypothesis testing
5. Annotator reliability and representativeness
6. Post-hoc tests
7. Practical coding session
8. Unit summary
9. References

Connection to pre-registration & pre-analysis

- A basic assumption of the statistical methods is that the hypothesis, test, sample sizes are fixed **before** an analysis
- "HARKing" (Hypothesizing After the Results are Known)
  - Basically, the after analysis hypothesis creation needs to be “corrected” for all other alternative hypotheses you may
  - Could be used for “exploratory” research, but NOT for ”confirmatory” research
- Pre-registration is the formalization of this act
- An extremely important choice is the number of samples N
- Power analysis estimates the required number of samples given an expected difference in values and variance

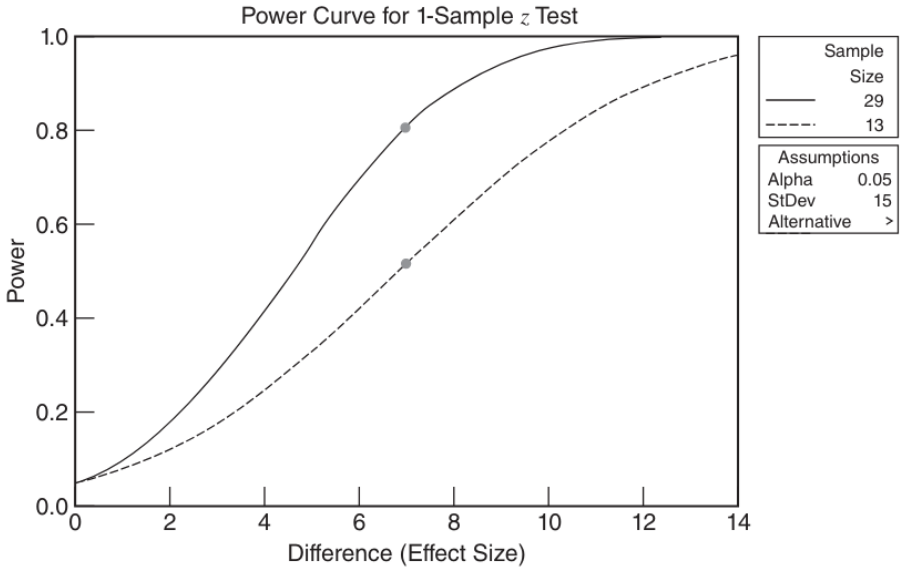
Power Analysis

- How much data do I need given an effect size?



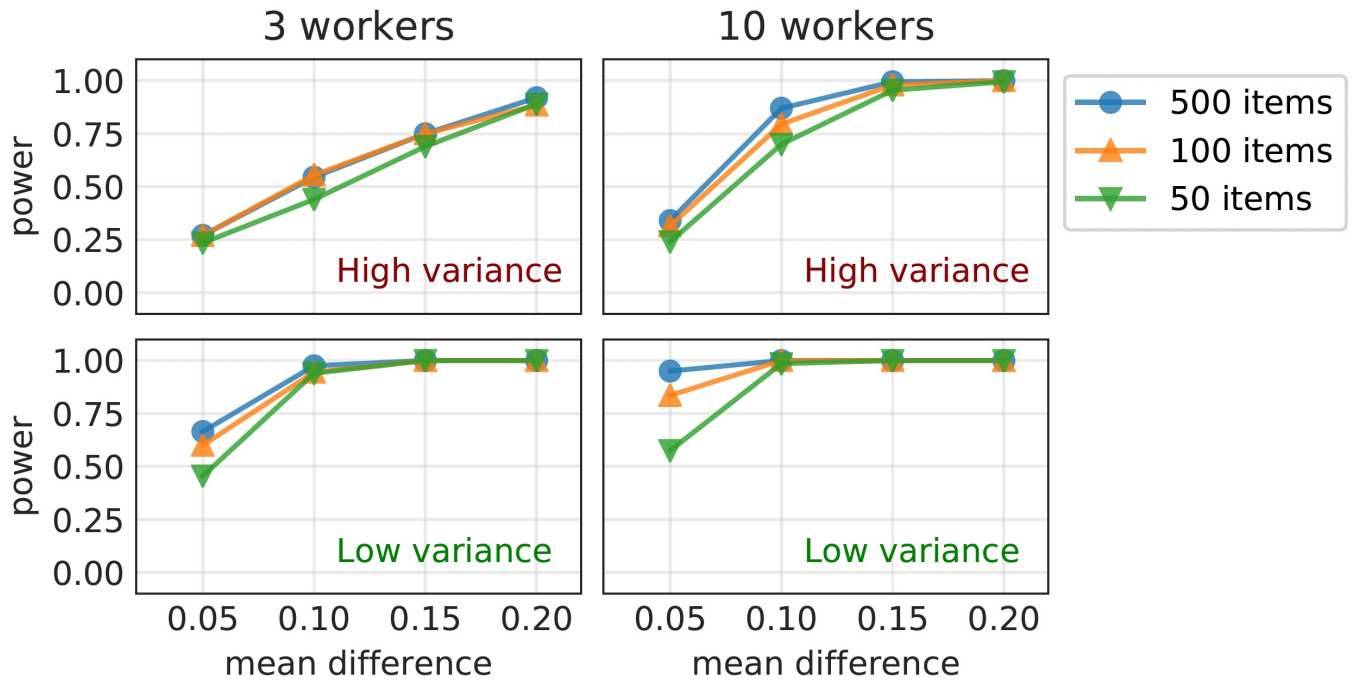
Power Analysis - Power Curves

- How much data do I need given an effect size?



Statistics, Whitte & Whitte, pg 232
https://nibmehub.com/opac-service/pdf/read/Statistics%20by%20Robert%20S.%20Witte-%20John%20S.%20Witte.pdf

Power Analysis



Source: With Little Power Comes Great Responsibility

Lots of NLP is paired (matched)

- Output of system A and B are based on the same input
- Paired t-test / McNemar / Cochran’s Q
- With pairing comes power

NHST: model reasoning VS outcome reasoning

This is fundamentally different

- Model building (“white box”)– “model reasoning”
  - There’s fundamental reason WHY your system A is better than the baseline and all other systems
- You are validating your new approach(es)

- Model ranking (“black box”) – “outcome reasoning”
- This comes from the Common Task Framework
- Fundamentally, we don’t care which approach is better
- Is there an order in which these are better?

- As always … there’s a “grey box” – multi-task CTF (understand the model through the task)

Source: When black box algorithms are (not) appropriate

Example 1: New method

I created a new type of language model for creative writing (“model reasoning”)

- My model will be better than vanilla LLM (baseline) as well as the current state-of-the-art system
  - During the writing process one of the 3 systems are presented to the user as a suggestion
  - Suggestions only happen once per interaction
  - Outcomes (judgments of usability on scale 5 point Likert scale)

- Properties:
  - Ordinal data
  - Items are independent (not matched)
  - Data is not normally distributed
  - Null hypothesis – there’s no difference between systems

Example 1: New method

I created a new type of language model for creative writing (“model reasoning”)

- My model will be better than vanilla LLM (baseline) as well as the current state-of-the-art system
  - During the writing process one of the 3 systems are presented to the user as a suggestion
  - Suggestions only happen once per interaction
  - Outcomes (judgments of usability on scale 5 point Likert scale)

- Properties:
  - Ordinal data
  - Items are independent (not matched)
  - Data is not normally distributed
  - Null hypothesis – there’s no difference between systems
  - H_A : new system is better than the other systems

Example 2: Shared Task with submitted systems

I want to rank system(s) for creative writing (“outcome reasoning”)

- In all cases we put the system in the SAME context and have annotator rate the system performance on a Likert scale for Fluency

- Properties:
  - Matched sample (item level dependency)
  - Desired outcome is a set of superior systems
  - Null Hypothesis – there’s no difference at all between the systems
  - Need to compute all pairwise rankings and correct for this

Example 3: Exploratory analysis

I want to try to create a hypothesis based on the outcome of the system(s) for creative writing

- I can’t do confirmatory analysis

- Properties:
  - Null Hypothesis – there’s effect of property X_1, …, X_n on the difference between the systems
  - Need multiple hypothesis testing correction
  - Conclusions are weaker

Overview

Unit 5: Analysis of Results

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Introduction to statistics – hypothesis testing
3. Connection to pre-registration (statistical power analysis)
4. Multiple hypothesis testing
5. Annotator reliability and representativeness
6. Post-hoc tests
7. Practical coding session
8. Unit summary
9. References

Multiple Hypothesis Testing

- False discovery rate

- Bonferroni correction
  Each individual hypothesis at a significance level of α/m, where α is the desired overall alpha level and m is the number of hypotheses

- Benjamini–Hochberg

  For a given alpha (α), find the largest k such that $P(k) \leq \frac{k}{m}\alpha$.

  Reject the null hypothesis (i.e., declare discoveries) for all $H(i)$ where $i = 1, \ldots, k$.

Overview

Unit 5: Analysis of Results

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Introduction to statistics – hypothesis testing
3. Connection to pre-registration (statistical power analysis)
4. Multiple hypothesis testing
5. Annotator reliability and representativeness
6. Post-hoc tests
7. Practical coding session
8. Unit summary
9. References

What is annotator agreement?

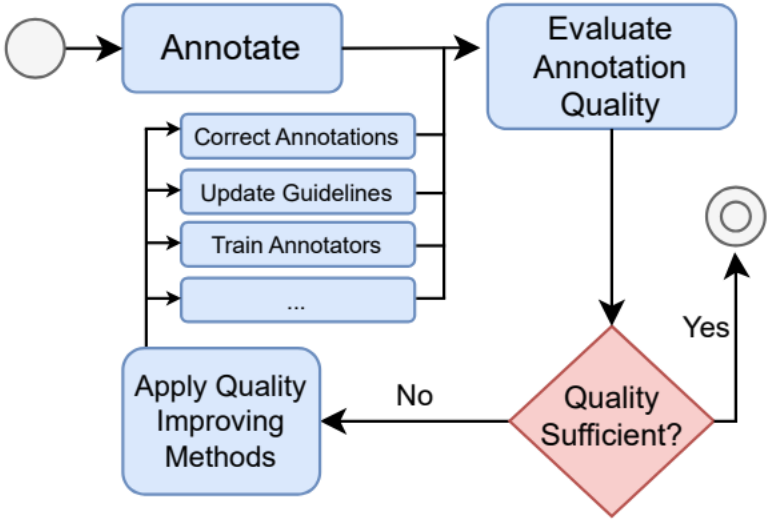


Figure 1: Overview of agile data corpus creation, the recommended workflow to annotate high-quality datasets. This work explores how to efficiently estimate annotation quality using statistics.

Source: On Efficient and Statistical Quality Estimation for Data Annotation

The Cohen’s Kappa metric is calculated as

$$\kappa = \frac{\text{observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}}$$

Source: **The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements**

IAA – Contingency Matrix

Rater 1

Rater 2

	17	8
	6	19

Source: https://datatab.net/tutorial/cohens-kappa

IAA – Cohen's Kappa

The Cohen's Kappa metric is calculated as follows:

$$(1)\quad \kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$$

where w_{ij}, x_{ij}, m_{ij} are the weight matrix, observed frequency matrix, and expected frequency matrix respectively.

$$(1)\quad \kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$$

where w_{ij}, x_{ij}, m_{ij} are the weight matrix, observed frequency matrix, and expected frequency matrix respectively.

Step 1: Create Weight Matrix

Weight	1	2	3	4	5
5	4	3	2	1	0
4	3	2	1	0	1
3	2	1	0	1	2
2	1	0	1	2	3
1	0	1	2	3	4

Step 2: Create Observed Frequency Matrix

Observed	1	2	3	4	5	RowSum
5	0	2	1	6	14	23
4	0	1	2	3	1	7
3	0	2	3	0	0	5
2	0	8	1	0	0	9
1	4	1	1	0	0	6
ColSum	4	14	8	9	15	**50**

Sum of rows

Step 1: Create Weight Matrix

Weight	1	2	3	4	5
5	4	3	2	1	0
4	3	2	1	0	1
3	2	1	0	1	2
2	1	0	1	2	3
1	0	1	2	3	4

Step 2: Create Observed Frequency Matrix

Observed	1	2	3	4	5	RowSum
5	0	2	1	6	14	23
4	0	1	2	3	1	7
3	0	2	3	0	0	5
2	0	8	1	0	0	9
1	4	1	1	0	0	6
ColSum	4	14	8	9	15	50

Sum of rows

Sum of columns

Step 3: Create Expected Frequency Matrix

The expected frequency at i,j:

(2) $$(p_e)_{i,j} = \frac{\sum_{i=1}^{k} x_{i,j} + \sum_{j=1}^{k} x_{i,j}}{n}$$

n = number of observations.

e.g.,

$$(p_e)_{3,2} = \frac{\sum_{i=1}^{k} x_{i,2} + \sum_{j=1}^{k} x_{3,j}}{50} = \frac{14 \times 5}{50} = 1.4$$

Expected	1	2	3	4	5
5	1.84	6.44	3.68	4.14	6.9
4	0.56	1.96	1.12	1.26	2.1
3	0.4	1.4 (14 * 5) / 50	0.8	0.9	1.5
2	0.72	2.52	1.44	1.62	2.7
1	0.48	1.68	0.96	1.08	1.8

Weight	1	2	3	4	5
5	4	3	2	1	0
4	3	2	1	0	1
3	2	1	0	1	2
2	1	0	1	2	3
1	0	1	2	3	4

Step 3: Create Expected Frequency Matrix

The expected frequency at i,j:

(2) $$(p_e)_{i,j} = \frac{\sum_{i=1}^{k} x_{i,j} + \sum_{j=1}^{k} x_{i,j}}{n}$$

n = number of observations.

e.g.,

$$(p_e)_{3,2} = \frac{\sum_{i=1}^{k} x_{i,2} + \sum_{j=1}^{k} x_{3,j}}{50} = \frac{14 \times 5}{50} = 1.4$$

Expected	1	2	3	4	5
5	1.84	6.44	3.68	4.14	6.9
4	0.56	1.96	1.12	1.26	2.1
3	0.4	1.4 (14 * 5) / 50	0.8	0.9	1.5
2	0.72	2.52	1.44	1.62	2.7
1	0.48	1.68	0.96	1.08	1.8

Step 4: Calculate the Weighted Cohen's Kappa

Observed Weighted	1	2	3	4	5
5	0	6	2	6	0
4	0	2	2	0	1
3	0	2	0	0	0
2	0	0	1	0	0
1	0	1	2	0	0
Sum	25				

Expected Weighted	1	2	3	4	5
5	7.36	19.32	7.36	4.14	0
4	1.68	3.92	1.12	0	2.1
3	0.8	1.4	0	0.9	3
2	0.72	0	1.44	3.24	8.1
1	0	1.68	1.92	3.24	7.2
Sum	80.64				

Step 4: Calculate the Weighted Cohen's Kappa

Observed Weighted	1	2	3	4	5
5	0	6	2	6	0
4	0	2	2	0	1
3	0	2	0	0	0
2	0	0	1	0	0
1	0	1	2	0	0
Sum	25				

Expected Weighted	1	2	3	4	5
5	7.36	19.32	7.36	4.14	0
4	1.68	3.92	1.12	0	2.1
3	0.8	1.4	0	0.9	3
2	0.72	0	1.44	3.24	8.1
1	0	1.68	1.92	3.24	7.2
Sum	80.64				

Using $^{(1)}$:

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}} = 1 - \frac{25}{80.64} \approx 0.69$$

IAA – What is the right level of alpha?

- Best method – researchers to pilot as experts and use consensus agreement or disagreement about items
- Established method – look at similar datasets and see the level of agreement
- Default method – use the recommended ranges

Value of κ	Strength of Agreement
< 0.20	Poor
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Good
> 0.80	Very Good

Source: https://datatab.net/tutorial/cohens-kappa

IAA

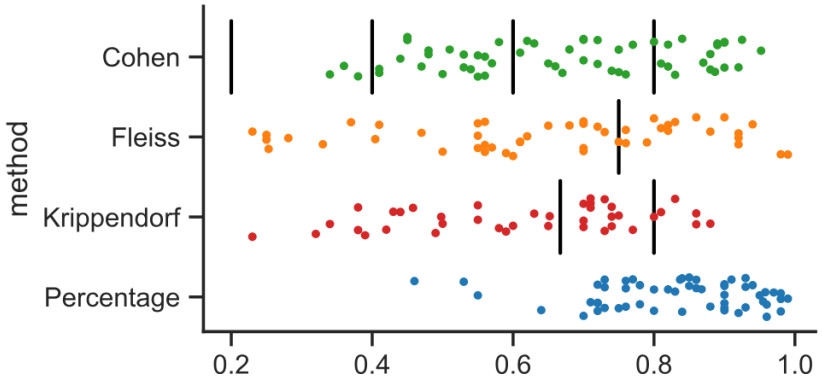


Figure 9: Agreement values for the papers inspected. Also shown are the ranges often used for interpreting these values.

Coefficient	# used	Average	Min.	Max.
Percent agreement	7	0.69	0.44	0.94
Cohen's κ	4	0.40	0.10	0.88
Krippendorff's α	5	0.62	0.37	0.90
Fleiss's κ	5	0.53	0.29	0.78
Pearson's r	2	0.42	0.20	0.71
Kendall's W	1	0.61	0.47	0.76
Weighted κ	1	0.07	0.07	0.07
κ no better specified	3	0.57	0.32	0.77

Table 3: Average, minimal and maximum IAA value per coefficient. *# used* means the number of times that a coefficient was used in total across the papers. In each paper each coefficient was used to measure the annotator's agreement about one or more questions or criteria.

Analyzing Dataset Annotation Quality Management in the Wild
Agreement is overrated: A plea for correlation to assess human evaluation reliability

IAA

- Fundamental misconceptions : you can make reliable conclusions without reliable agreement (but … your power is lower)
- Descriptive VS prescriptive items aren't differentiated in NLP
  - Descriptive paradigm **encourages** annotator subjectivity to create datasets as granular surveys of individual beliefs
  - Prescriptive paradigm **discourages** annotator subjectivity and instead tasks annotators with encoding one specific belief, formulated in the annotation guidelines
- There's completely different ideas
  - Test-retest reliability
  - Multi-item testing Cronback's alpha
- There's other correlation metrics
  - Inter-class correlations
  - Goodman and Kruskal's Gamma
- What about a simple contingency table? Agreement %?
- Model based validity – can I predict disjoint out of sample data?

Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks
Agreement is overrated: A plea for correlation to assess human evaluation reliability

Overview

Unit 5: Analysis of Results

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Introduction to statistics – hypothesis testing
3. Connection to pre-registration (statistical power analysis)
4. Multiple hypothesis testing
5. Annotator reliability and representativeness
6. Post-hoc tests
7. Practical coding session
8. Unit summary
9. References

Post-hoc Tests

- Correlations between outcomes and other variables
- Analysis of variance (ANOVA) or Kruskal-Wallis
- MUST at least be corrected for multiple tests
- These should be pre-registered is possible
- If not pre-registered then these should in no way be thought of as confirmatory

Overview

Unit 5: Analysis of Results

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Introduction to statistics – hypothesis testing
3. Connection to pre-registration (statistical power analysis)
4. Multiple hypothesis testing
5. Annotator reliability and representativeness
6. Post-hoc tests
7. Practical coding session
8. Unit summary
9. References

Unit Summary

- Discuss data transformations and model-free evidence
- Analysis of results from a statistical perspective
- Null hypothesis significance testing, statistical significance tests, and power analysis
- Gain an understanding of pre-registration and confirmatory vs. exploratory hypothesis testing
- Understand annotator agreement metrics
- Understand post-hoc analysis and multiple hypothesis testing corrections for false discovery

References

Essential References

- With Little Power Comes Great Responsibility
- Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks
- The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing
- Improving the Science of Annotation for Natural Language Processing: The Use of the Single-Case Study for Piloting Annotation Projects

- Interrater Disagreement Resolution: A Systematic Procedure to Reach Consensus in Annotation Tasks

Extended References

https://digital.lib.washington.edu/server/api/core/bitstreams/ead01b90-c7a1-448f-86be-612bf422c9e8/content
https://link.springer.com/book/10.1007/978-3-031-02174-9
https://aclanthology.org/W14-1601.pdf (p-value in NLP)
"**Power failure: why small sample size undermines the reliability of neuroscience**"
With Little Power Comes Great Responsibility
Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155–159
The Authenticity Gap in Human Evaluation https://arxiv.org/pdf/2205.11930
Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks https://aclanthology.org/2022.naacl-main.13.pdf

Interrater Disagreement Resolution: A Systematic Procedure to Reach Consensus in Annotation Tasks
Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines
Best practices for the human evaluation of automatically generated text
On Efficient and Statistical Quality Estimation for Data Annotation

Evaluating Human Pairwise Preference Judgments https://aclanthology.org/J15-2005.pdf - Bradley-Terry Appraoch
IRT - . Models of translation competitions – Hopkins & May
Trueskill - Efficient Elicitation of Annotations for Human Evaluation of Machine Translation

Redefine statistical significance https://pubmed.ncbi.nlm.nih.gov/30980045/
Three Recommendations for Improving the Use of *p*-Values
Abandon Statistical Significance https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1527253#abstract https://arxiv.org/pdf/1709.07588
Beyond p values: utilizing multiple methods to evaluate evidence https://link.springer.com/article/10.1007/s41237-019-00078-4
Moving to a World Beyond "p < 0.05" https://www.tandfonline.com/doi/pdf/10.1080/00031305.2019.1583913

Extended References

- Hard and Soft Evaluation of NLP models with BOOtSTrap SAmpling – BooStSa (really automatic metrics)
- Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation (should we do error based analysis)
- *Eras: Improving the quality control in the annotation process for Natural Language Processing*
  tasks https://www.sciencedirect.com/science/article/abs/pii/S0306437920300521
- Improving the Science of Annotation for Natural Language Processing: The Use of the Single-Case Study for Piloting Annotation Projects
- Hierarchical Evaluation Framework: Best Practices for Human Evaluation

References (for learning)

- **Statistical Significance Testing for Natural Language Processing**

- Mathematical Statistics: Basic Ideas and Selected Topics
- All of Statistics: A Concise Course in Statistical Inference

- **Cracking the Code of Data: A Student's Guide to Hypothesis Testing**
  https://www.quanthub.com/cracking-the-code-of-data-a-students-guide-to-hypothesis-testing/

Lots More I Couldn’t Cover

Sorry … the story is more complicated

- In statistical testing people are no longer satisfied with hypothesis testing.
- Other methods like Bayes factors, Observation oriented modeling, Akaike Information Criterion (AIC)
- Basic philosophy – never trust a single number
  - System A > System B
    - Why?
    - Does it replicate to other populations?
    - Does it replicate to other datasets?
    - Is it only superior on a subset of the data?

- Add correlation between human and automatic metrics
- Show chance correction

Hypothesis Testing

What to say instead –

Give example

DON’T SAY “STATISTICALLY SIGNIFICANT” –

Edgeworth’s (1885) original intention for statistical significance was simply as a tool to indicate when a result warrants further scrutiny

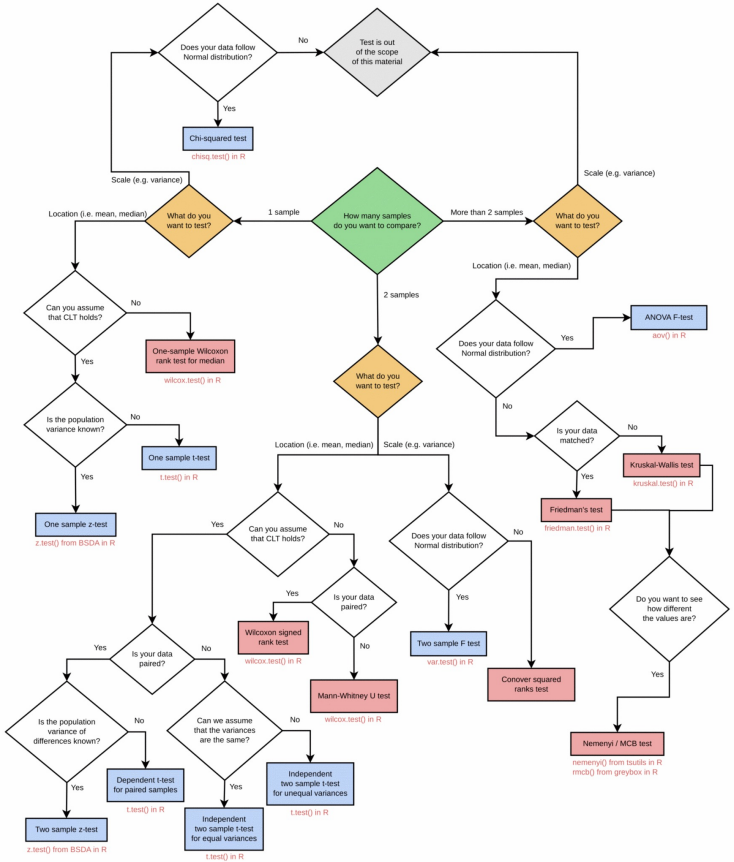
Gelman and Stern (2006) famously observed, the difference between “significant” and “not significant” is not itself statistically significant

Source: https://www.youtube.com/watch?v=uHUOwZmfUzY

Picking the rig

Flowchart for selection of a statistical test about location / scale

To select the test, start from the green rhombus



Developed by Ivan Svetunkov, CMAF, Lancaster University: https://forecasting.svetunkov.ru/
Read more about this at https://openforecast.org/sba/

Power Analysis – oh there's more …

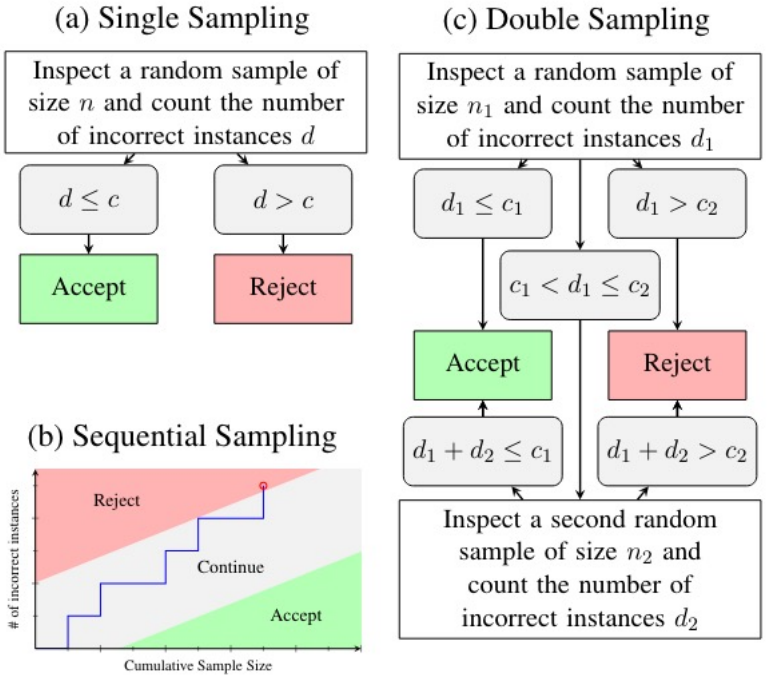
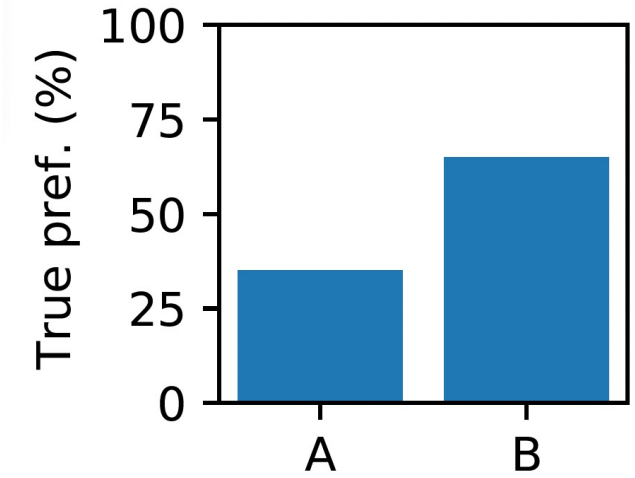
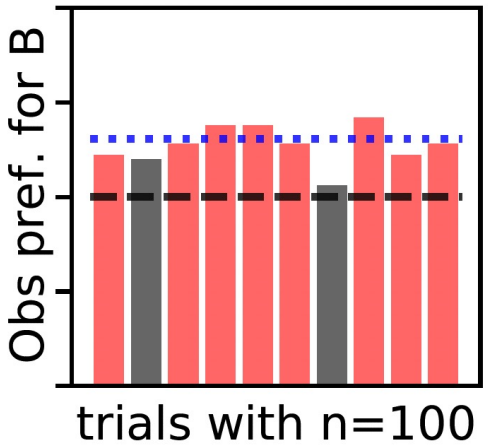


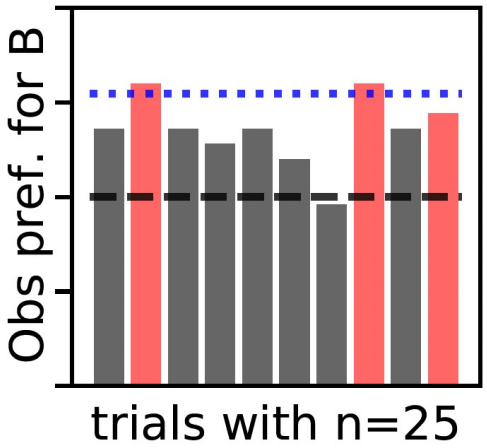
Figure 2: Flowcharts for the three different acceptance sampling methods discussed in this work.

Source: On Efficient and Statistical Quality Estimation for Data Annotation

Power Analysis

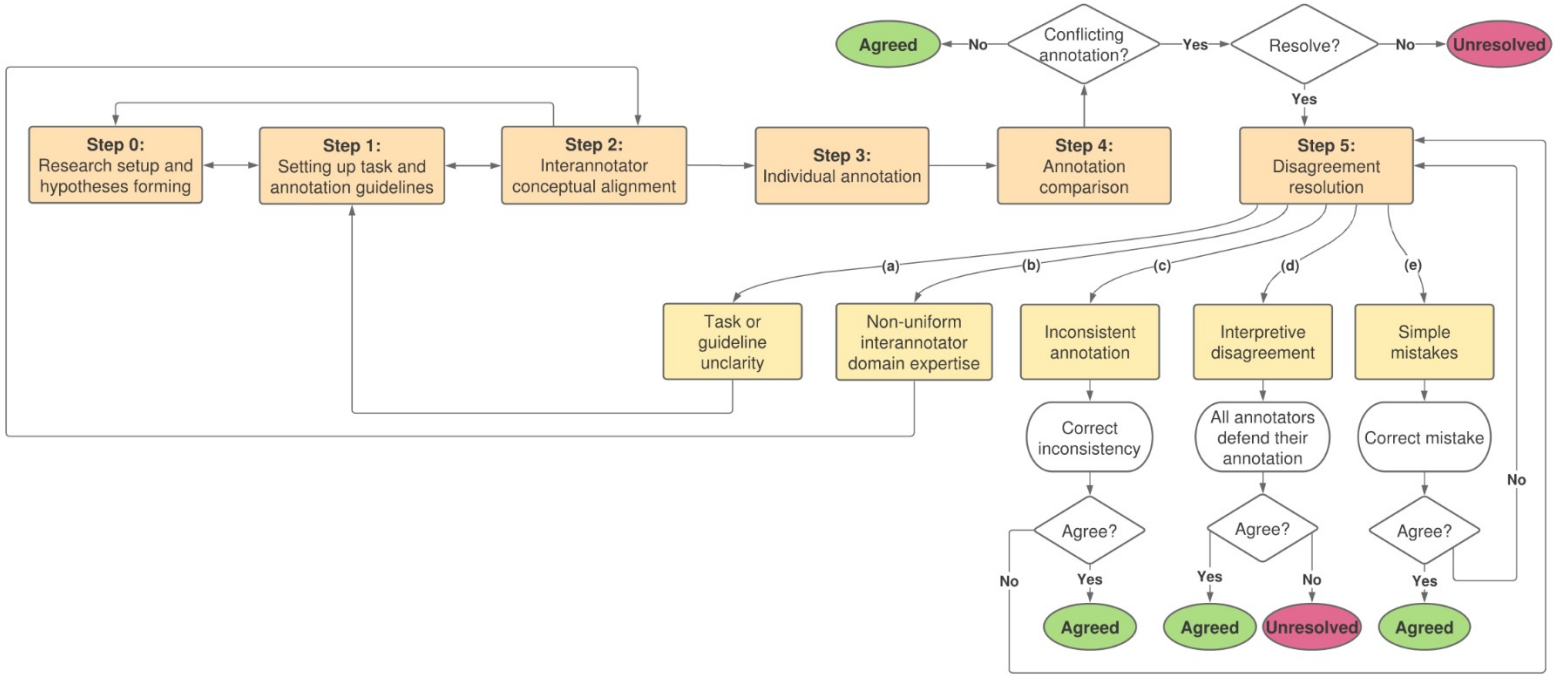






Source: [With Little Power Comes Great Responsibility]

What is annotator agreement?



Source: Interrater Disagreement Resolution: A Systematic Procedure to Reach Consensus in Annotation Tasks

What is annotator agreement?

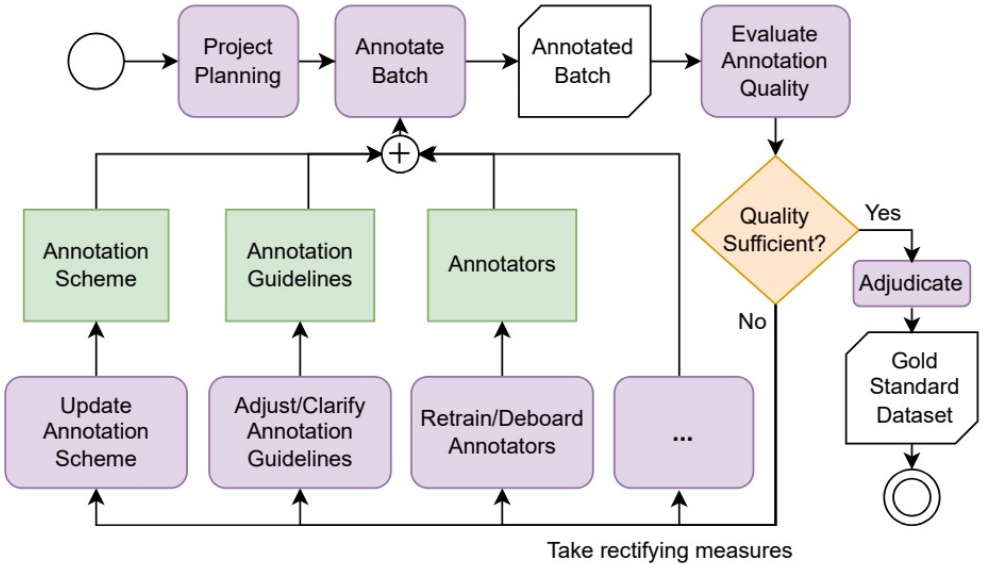


Figure 2: The recommended annotation process: After a batch of data is annotated, it is evaluated. If the quality is sufficient, it can be adjudicated. If not, several corrective measures can be taken, e.g., correcting the annotations in an additional step, annotator training, or adjusting the annotation scheme or guidelines. This is similarly applicable for text production workflows where usually no adjudication takes place.

Source: Analyzing Dataset Annotation Quality Management in the Wild

IAA

Figure 1
Generic 2-by-2 Contingency Table

Values:		Coder A 0	Coder A 1	
Coder B	0	a	b	p_B
	1	c	d	q_B
		p_A	q_A	1

Population Estimates

$\bar{p} = (p_A + p_B)/2$ = proportion of 0s in data

$\bar{q} = (q_A + q_B)/2 = 1 - \bar{p}$ = proportion of 1s in data

Figure 2 states the above-mentioned agreement coefficients in terms of Figure 1 and in α's economical form:

$$\text{Agreement} = 1 - \frac{D_o}{D_e} = 1 - \frac{\text{Observed Disagreement}}{\text{Expected Disagreement}}$$

where, when the observed disagreement D_o=0, agreement =1; and when the two disagreements are equal, D_o=D_e, agreement =0. So, D_o expresses the lack of agreement; whereas D_e defines the zero point of the measure.

Figure 2
The Dichotomous Forms of Seven Agreement Coefficients

	Agreement	=	**1**	–	**Observed**	/	**Expected Disagreements**	
%-agreement	$\mathbf{A_o}$	=	1	–	(b + c)			
Osgood (1959); Holsti's	**CR**	=	$[1 - (b+c)]$		$\frac{2 \cdot N_{A \cap B}}{N_A + N_B}$			
Bennett et al. (1954)	**S**	=	1	–	(b + c)	/	$2 \cdot ½ \cdot ½$	where ½ is the logical probability of 0 and of 1
Scott (1955)	**π**	=	1	–	(b + c)	/	$2\bar{p}\bar{q}$	where $\bar{p} = \frac{p_A + p_B}{2}$ and $\bar{q} = 1 - \bar{p}$
Krippendorff (1970)	**α**	=	1	–	(b + c)	/	$\frac{n}{n-1} 2\bar{p}\bar{q}$	where n = the number of 0s and 1s used jointly
Cohen (1960)	**κ**	=	1	–	(b + c)	/	$p_A q_B + p_B q_A$	

IAA – Alpha hacking

- Don’t do it!
- Post fact removal of annotators ☹
  - MACE – sorry – Zhang et al., 2023 showed that post-hoc statistical filtering may not agree with filtered reliable annotators

A/B testing VS Multi-armed Bandit

- Learning from a payoff for being correct
- Rather than 50/50 A/B split dynamically allocate the systems
- Algorithm can be done online
- Currently, almost never done in NLP research
- Potentially, this could be tested with off-policy learning

Source: Machine Translation System Selection from Bandit Feedback