Inter-Annotator Agreement II

Multiple Annotators and Advanced Measures

Jin Zhao

Brandeis University

March 25, 2025

Today's Agenda

1. Review of Cohen's Kappa
2. Fleiss' Kappa (multiple annotators)
3. Krippendorff's Alpha
4. Agreement for sequence labeling (spans)
5. Agreement for rankings
6. Human-LLM agreement
7. LLM self-consistency

Review: Cohen's Kappa

For two annotators:

$$\kappa = \frac{A_o - A_e}{1 - A_e}$$

Limitations:

- Only works for exactly 2 annotators
- Assumes same 2 annotators for all items
- Categorical labels only

Today: Extensions for more complex scenarios

Fleiss' Kappa: Multiple Annotators

When you have 3+ annotators

Same formula structure:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Key differences from Cohen's:

- Works with any number of annotators
- Different annotators can label different items
- More complex calculation of expected agreement

Assumption: Fixed number of annotators per item (e.g., always 3)

Fleiss' Kappa Calculation

For n items, k categories, r annotators per item:

Let n_{ij} = number of annotators who assigned item i to category j

Per-item agreement:

$$P_i = \frac{1}{r(r-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$

Mean observed agreement:

$$\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i$$

Expected agreement:

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2$$

where p_j = proportion of all ratings in category j

Fleiss' Kappa Example

10 items, 3 categories, 4 annotators each

Item	Cat A	Cat B	Cat C	P_i
1	4	0	0	1.00
2	3	1	0	0.50
3	2	2	0	0.33
4	0	4	0	1.00
5	1	2	1	0.17
Totals	...	...	...	

Use Python: `statsmodels.stats.inter_rater.fleiss_kappa`
Or `nltk.metrics.agreement`

Krippendorff's Alpha

Most flexible agreement measure

Advantages:

- Any number of annotators
- Missing data allowed
- Works with different data types:
  - Nominal (categories)
  - Ordinal (rankings)
  - Interval (ratings)
  - Ratio (measurements)

$$\alpha = 1 - \frac{D_o}{D_e}$$

Where D_o = observed disagreement, D_e = expected disagreement

Krippendorff's Alpha: Key Features

Different distance functions:

- **Nominal:** $d^2 = 0$ if same, 1 if different
- **Ordinal:** d^2 based on rank distance
- **Interval:** $d^2 = (c - k)^2$
- **Ratio:** $d^2 = \frac{(c-k)^2}{(c+k)^2}$

Interpretation:

- $\alpha = 1$: Perfect agreement
- $\alpha = 0$: Agreement equals chance
- $\alpha < 0$: Systematic disagreement

Guideline: $\alpha > 0.8$ reliable, $\alpha > 0.67$ acceptable

Agreement for Sequence Labeling

NER and span annotation require special measures

Two approaches:

1. **Token-level:** Treat each token as a classification
   - Use standard Kappa on BIO labels
   - Doesn't capture span-level structure
2. **Entity-level:** Compare extracted spans
   - Precision/Recall/F1 between annotators
   - Exact match vs. partial match

Entity-Level Agreement

Comparing annotator spans:

Exact match:

- Spans must have identical boundaries AND type
- Strict but clear

Partial match (relaxed):

- Allow some boundary variation
- More forgiving of minor differences

Metrics:

$$P = \frac{|A \cap B|}{|A|}, \quad R = \frac{|A \cap B|}{|B|}, \quad F_1 = \frac{2PR}{P + R}$$

Where A = spans from annotator 1, B = spans from annotator 2

Agreement for Rankings

For preference annotation:

Pairwise agreement:

- Do annotators agree on which is better?
- Simple percentage or Kappa

Rank correlation:

- **Spearman's ρ:** Correlation of rank positions
- **Kendall's τ:** Proportion of concordant pairs

For RLHF:

- Often use simple majority vote
- Low agreement may be acceptable (captures preference diversity)

Human-LLM Agreement

New measure: How well does LLM match human annotation?

Calculate same as human-human:

- LLM as "annotator 2"
- Compute Kappa, F1, etc.

What to measure:

- LLM vs. single human annotator
- LLM vs. gold standard (adjudicated)
- LLM vs. majority vote

Interpretation:

- High agreement: LLM suitable for task
- Low agreement: Need human annotation

LLM Self-Consistency

Does the LLM agree with itself?

Method:

1. Run same prompt multiple times
2. Compare outputs across runs
3. Calculate agreement metrics

Why it matters:

- High consistency: Reliable (though not necessarily correct)
- Low consistency: Unreliable, needs human review

Use for confidence estimation:

- If LLM gives same answer 10/10 times: high confidence
- If LLM varies across runs: route to human

Choosing the Right Measure

Scenario	Recommended Measure
2 annotators, categories	Cohen's Kappa
3+ annotators, categories	Fleiss' Kappa
Variable annotators, categories	Krippendorff's Alpha
Ordinal ratings	Krippendorff's Alpha (ordinal)
Sequence labeling	Token Kappa + Entity F1
Rankings	Kendall's τ
Human-LLM comparison	Same as human-human

Python Tools for IAA

Libraries:

- `sklearn.metrics.cohen_kappa_score` – Cohen's Kappa
- `statsmodels.stats.inter_rater` – Fleiss' Kappa
- `krippendorff` – Krippendorff's Alpha (pip install)
- `nltk.metrics.agreement` – Multiple measures
- `scipy.stats.kendalltau` – Rank correlation

NLTK AnnotationTask:

- Flexible input format
- Multiple agreement metrics
- Handles missing data

Reporting Multi-Annotator Agreement

What to include:

1. Number of annotators
2. Number of items
3. Agreement metric used (and why)
4. Agreement value with interpretation
5. Per-category breakdown if applicable
6. Comparison to baselines or prior work

Example:

"Three annotators labeled 200 sentences. Fleiss' Kappa was 0.68, indicating substantial agreement. Per-category F1: PER=0.85, ORG=0.72, LOC=0.78."

Next Class: IAA & Modeling Introduction

Lecture 20 (Mar 30): IAA Wrap-up — Modeling Introduction

Note: No class April 1 (Passover)

Topics:

- Resolving annotator disagreements
- Adjudication strategies
- Creating gold standard datasets
- LLM-assisted adjudication
- Introduction to modeling with annotated data

Assignment: HW 3 due

Key Takeaways

1. **Fleiss' Kappa** extends Cohen's to multiple annotators
2. **Krippendorff's Alpha** is most flexible (any data type, missing data)
3. **Span agreement** needs both token and entity-level measures
4. **Ranking agreement** uses correlation measures
5. **Human-LLM agreement** validates LLM annotation quality
6. **LLM self-consistency** can indicate confidence

Questions?

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

jinzhao@brandeis.edu