Inter-Annotator Agreement I

Cohen's Kappa and Basic Agreement Measures

Jin Zhao

Brandeis University

March 23, 2025

Today's Agenda

1. Why measure agreement?
2. Observed agreement
3. Problems with raw agreement
4. Cohen's Kappa (2 annotators)
5. Calculating Kappa step by step
6. Interpreting Kappa values
7. Worked examples

Project: IAA evaluation due this week
Assignment: HW 3 assigned

Why Measure Agreement?

IAA serves multiple purposes:

1. **Quality assessment:** How reliable is our data?
2. **Task validation:** Is the task well-defined?
3. **Guideline evaluation:** Are guidelines clear?
4. **Annotator calibration:** Are annotators consistent?
5. **Upper bound:** What's the best model can achieve?

Key insight: If humans can't agree, models can't learn reliably

Observed Agreement

Simplest measure: Proportion of items where annotators agree

$$A_o = \frac{\text{Number of agreements}}{\text{Total items}}$$

Example:

Item	Ann. 1	Ann. 2	Agree?
1	Positive	Positive	✓
2	Negative	Negative	✓
3	Positive	Neutral	×
4	Negative	Negative	✓
5	Neutral	Neutral	✓

$$A_o = \frac{4}{5} = 0.80$$

Problem: Chance Agreement

Observed agreement doesn't account for chance

Example: Binary task (Yes/No)

- Two annotators randomly guessing
- Each says "Yes" 50% of the time
- Expected agreement by chance: 50%

If we observe 70% agreement:

- Seems good, but...
- Only 20% above chance
- Is that meaningful?

Solution: Chance-corrected measures like Kappa

Cohen's Kappa: The Formula

Kappa corrects for chance agreement

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

Where:

- A_o = Observed agreement
- A_e = Expected agreement by chance
- κ = Kappa coefficient

Intuition: Agreement beyond chance, normalized by maximum possible agreement beyond chance

Computing Expected Agreement

A_e based on marginal distributions

Confusion matrix:

	Ann. 2: Yes	Ann. 2: No	Total
Ann. 1: Yes	a	b	$a+b$
Ann. 1: No	c	d	$c+d$
Total	$a+c$	$b+d$	n

$$A_e = P(\text{both Yes}) + P(\text{both No})$$

$$A_e = \frac{(a+b)(a+c)}{n^2} + \frac{(c+d)(b+d)}{n^2}$$

Step-by-Step Kappa Calculation

Example: Sentiment annotation (Pos/Neg)

	Ann. 2: Pos	Ann. 2: Neg	Total
Ann. 1: Pos	20	5	25
Ann. 1: Neg	10	15	25
Total	30	20	50

Step 1: Observed agreement

$$A_o = \frac{20 + 15}{50} = 0.70$$

Step 2: Expected agreement

$$A_e = \frac{25 \times 30}{50^2} + \frac{25 \times 20}{50^2} = 0.30 + 0.20 = 0.50$$

Step 3: Kappa

$$\kappa = \frac{0.70 - 0.50}{1 - 0.50} = \frac{0.20}{0.50} = 0.40$$

Interpreting Kappa Values

Common interpretation guidelines (Landis & Koch, 1977):

Kappa	Interpretation
< 0	Less than chance agreement
$0.00 - 0.20$	Slight agreement
$0.21 - 0.40$	Fair agreement
$0.41 - 0.60$	Moderate agreement
$0.61 - 0.80$	Substantial agreement
$0.81 - 1.00$	Almost perfect agreement

Caution: These are rough guidelines, not absolute thresholds

What Kappa Should You Aim For?

Depends on the task:

- **Objective tasks** (POS tagging): $\kappa > 0.80$
- **Standard annotation** (NER, sentiment): $\kappa > 0.70$
- **Subjective tasks** (quality, safety): $\kappa > 0.50$
- **Highly subjective** (sarcasm, humor): $\kappa > 0.40$

Context matters:

- Compare to published baselines for similar tasks
- Consider task difficulty
- Low Kappa may indicate unclear guidelines, not bad annotators

When Kappa Can Be Misleading

Kappa paradoxes:

1. **High agreement, low Kappa:**
   - When one category dominates (class imbalance)
   - High A_e reduces Kappa
2. **Symmetric vs. asymmetric disagreement:**
   - Kappa treats all disagreements equally
   - Some disagreements may be more problematic
3. **Prevalence effect:**
   - Rare categories have outsized impact

Recommendation: Report Kappa AND confusion matrix

Kappa for Multi-Class

Same formula, larger confusion matrix

For k categories:

$$A_o = \frac{1}{n} \sum_{i=1}^{k} n_{ii}$$

$$A_e = \sum_{i=1}^{k} \frac{n_{i\cdot} \times n_{\cdot i}}{n^2}$$

Where n_{ii} is the count in cell (i, i), and $n_{i\cdot}$, $n_{\cdot i}$ are row and column totals.

Multi-Class Example

3-class sentiment: Positive / Neutral / Negative

	Pos	Neu	Neg	Total
Pos	30	5	5	40
Neu	3	20	2	25
Neg	2	3	30	35
Total	35	28	37	100

$A_o = \frac{30+20+30}{100} = 0.80$
$A_e = \frac{40 \times 35}{10000} + \frac{25 \times 28}{10000} + \frac{35 \times 37}{10000} = 0.339$
$\kappa = \frac{0.80-0.339}{1-0.339} = 0.697$ (Substantial agreement)

Using Python for Kappa

scikit-learn implementation:

```
from sklearn.metrics import cohen_kappa_score

ann1 = ['pos', 'neg', 'pos', 'neg', 'neu']
ann2 = ['pos', 'neg', 'neu', 'neg', 'neu']

kappa = cohen_kappa_score(ann1, ann2)
print(f"Kappa: {kappa:.3f}")
```

Output: `Kappa: 0.600`

Reporting Agreement

What to include in your report:

1. **Kappa value** with interpretation
2. **Observed agreement** for context
3. **Confusion matrix** for detailed view
4. **Per-class agreement** if relevant
5. **Sample size** (number of items annotated)
6. **Number of annotators**

Example: "Two annotators labeled 500 items for sentiment. Cohen's Kappa was 0.72 (substantial agreement) with observed agreement of 85%."

What If Agreement Is Low?

Diagnosis and remediation:

1. **Unclear guidelines:**
   - Add examples and edge cases
   - Clarify ambiguous definitions
2. **Insufficient training:**
   - More annotator calibration sessions
   - Practice on sample data
3. **Task too subjective:**
   - Simplify categories
   - Accept lower agreement threshold
4. **Bad annotators:**
   - Review individual performance
   - Remove or retrain outliers

Next Class: IAA II

Lecture 19 (Mar 25): Inter-Annotator Agreement II

Topics:

- Fleiss' Kappa (multiple annotators)
- Krippendorff's Alpha
- Agreement for spans and rankings
- Human-LLM agreement measurement
- LLM self-consistency as quality signal

Reading: Artstein (2017) - Handbook chapter on IAA

Key Takeaways

1. **IAA** measures annotation quality and task clarity
2. **Observed agreement** doesn't account for chance
3. **Cohen's Kappa** corrects for chance agreement
4. $\kappa = \frac{A_o - A_e}{1 - A_e}$ measures agreement beyond chance
5. **Interpretation** depends on task type
6. **Low agreement** may indicate guideline or task problems

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

jinzhao@brandeis.edu