Inter-Annotator Agreement Part 2

LIN 313 Language and Computers
UT Austin Fall 2025
Instructor: Gabriella Chronis

Admin 10/1

- Homework 2 was due monday
  - feedback posted early next week
  - thank you for the feedback!
    - study guides / unit summaries
    - 'follow along' activities + visual aides
    - too much reading / don't talk about the reading in class
    - how does gpt5 work?
- reading + annotation for Friday: "Excavating AI"
- homework 3 posted this afternoon - due Wednesday, October 15

ChatGPT minecraft implementation

Is it actually ChatGPT?

CraftGPT

- 5,087,280 parameters,
- **trained in Python**
- on the TinyChat dataset of basic English conversations.
- |embedding dimension| = 240
- |vocabulary| = 1920
- 6 layers
- The context window size is 64 tokens, (enough for (very) short conversations.
- Weights were quantized to 8 bits,
    - although the embedding and LayerNorm weights are stored at 18 and 24 bits respectively.

ChatGPT (all estimates bc undisclosed)

- 1.8 *trillion - 10 trillion* parameters
- **trained in Python on the**
- on an undisclosed dataset (probably the entire internet
- |embedding dimension| = ~16000
- |vocabulary| = 199,997
- 48 layers
- The context window size is 400,00 tokens
- Weight quantization = ??
    - this depends on hardware

Objectives

- writing successful annotation guidelines
- understanding the components of Cohen's Kappa
  - reasoning about how the terms relate to one another
  - able to apply the formula to calculate agreement on real examples

Writing good guidelines

We want our data to reflect the annotator's intuitions about the meaning of the category, not our (the researcher's) intuitions. In some cases, especially business applications, you may need to clearly define your categories.

In our case, we are interested in gathering data about speaker intuitions. Therefore,

We don't want to interpret the category or reduce the category to a related feature

e.g. label 'pick up line' if the sentence contains a joke

e.g. label 'pick up line' if the sentence is cringe and 'trash talk' if the sentence is based.

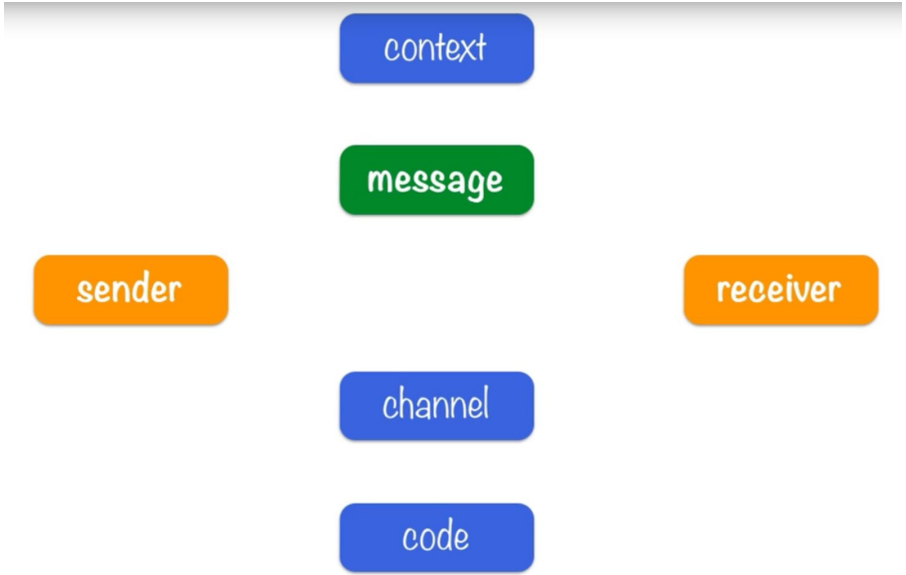
We do want to explain how to apply the category

e.g. label 'pick up line' if you would be more likely to use it to flirt

e.g label 'pick up line' if you would be more likely to think someone who said it is flirting

Writing good annotation guidelines

One way to write a good guideline is to ask: what *part* of the speech situation is this judgment applying to?

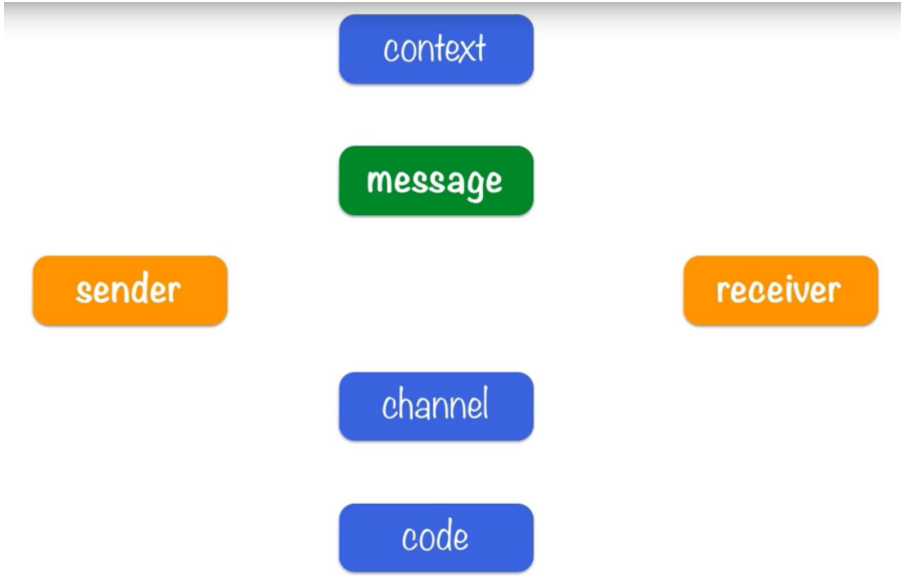


- code: choose 'pick up line' if this sentence uses the **type of language** used in pick up lines
- context: choose 'pick up line' if this sentence **describes a situation** more related to pick up lines
- receiver: choose pick up line if **hearing this would make you think** someone was interested in you

Practice writing annotation guidelines

The classification scheme:

- red flag 🚩🚩 vs. green flag ✅✅✅
- write a classification guideline that directs the annotator to reason about the **context** (a.k.a. the referential content of the sentence)

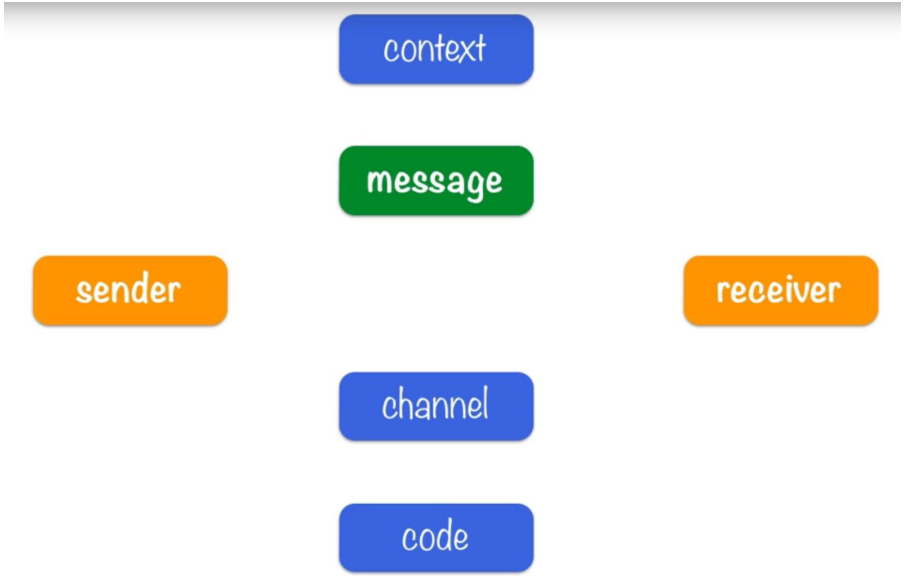


https://polls.la.utexas.edu/course/5422/teacher#/polls/10%20-%20Inter-annotator%20agreement%20II

More practice writing annotation guidelines

The classification scheme:

- red flag 🚩🚩 vs. green flag ✅✅✅
- write a classification guideline that directs the annotator to reason about the **message** (a.k.a. the formal stylistic characteristics of the sentence)



https://polls.la.utexas.edu/course/5422/teacher#/polls/10%20-%20Inter-annotator%20agreement%20II

Measuring inter-annotator agreement

On Monday we ran some pilot experiments with three variables

1. the classification scheme
2. the annotation guidelines
3. the text data to be annotated

We evaluated these experiments using Cohen's Kappa (Cohen's κ)

Cohen's Kappa is a measure of how well two annotators are aligned with each other beyond random chance.

Cohen's Kappa (Cohen's κ)

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

P_0 = relative observed agreement

P_e = probability of random agreement

measures how much agreement isn't just due to random chance.

let's work through calculating P_0 and P_e in an example

Example: Sooji and Gabriella annotate pick up line vs trash talk

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- they both annotate fifty examples
- they agree that 20 are pick up lines
- they agree that 15 are trash talk

Example: Sooji and Gabriella annotate pick up line vs trash talk

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- they both annotate fifty examples
- they agree that 20 are pick up lines
- they agree that 15 are trash talk

What is their **percent agreement**?

P_0 : relative observed agreement (aka percent agreement)

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

P_0 : relative observed agreement

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

The proportion of cases where the two raters actually agreed. You just count up all the times they gave the same rating and divide by the total number of cases.
(this is just percent agreement, expressed as a decimal rather than a percent)

P_0 : relative observed agreement

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

P_0 = # agree / total

The proportion of cases where the two raters actually agreed. You just count up all the times they gave the same rating and divide by the total number of cases.
(this is just percent agreement, expressed as a decimal rather than a percent)

P_0 : relative observed agreement

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

P_0 = # agree / total

= (20 + 15) / 50

= 35 / 50

= 7 / 10

= 0.7

$P_0 = 0.7$

The proportion of cases where the two raters actually agreed. You just count up all the times they gave the same rating and divide by the total number of cases.
(this is just percent agreement, expressed as a decimal rather than a percent)

Cohen's Kappa

this is just % agreement

P_0 = relative observed agreement

P_e = probability of random agreement

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

Cohen's Kappa

this is just % agreement

P_0 = relative observed agreement

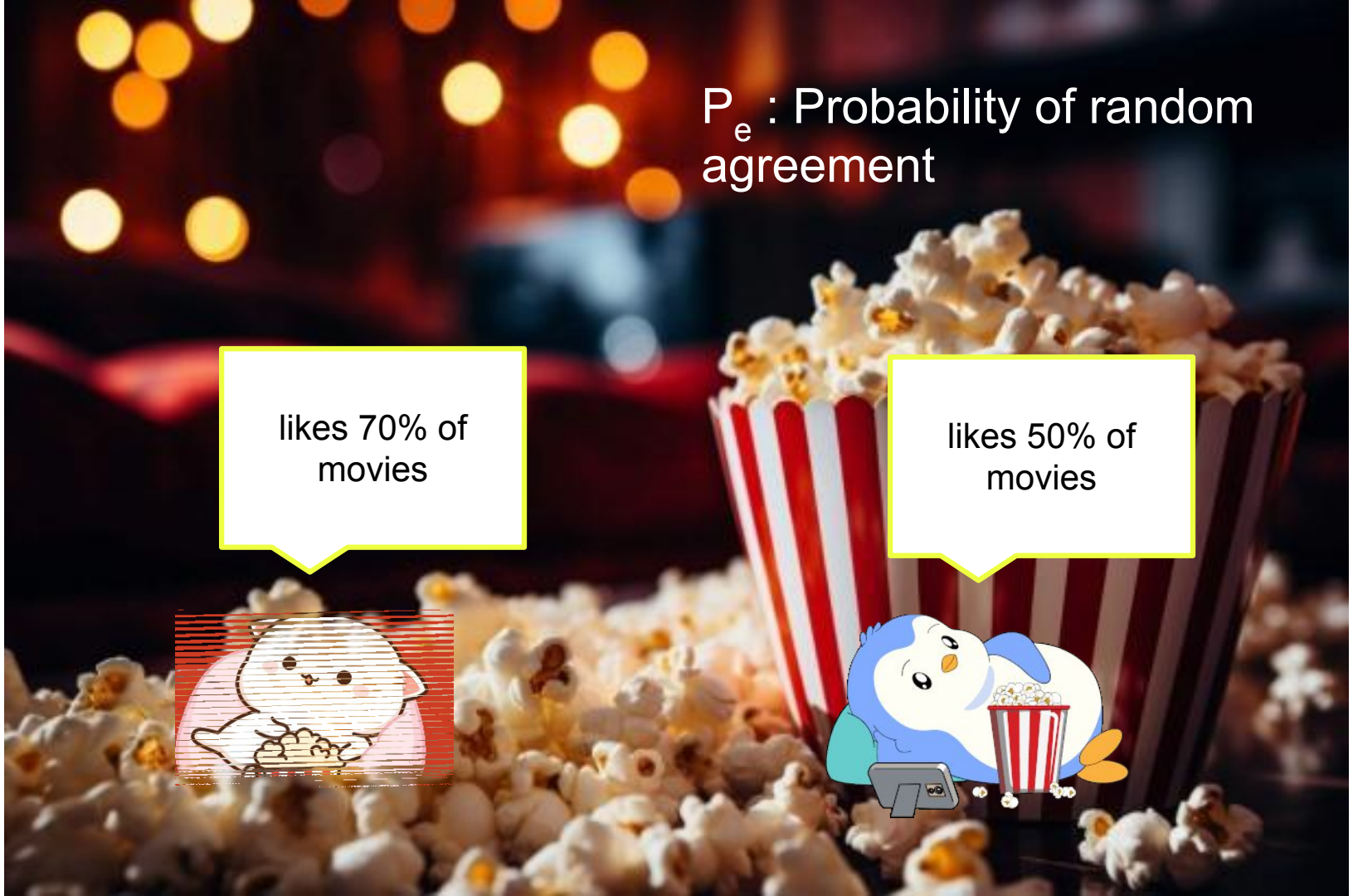
P_e = probability of random agreement

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

this one is trickier!


P_e : Probability of random agreement

P_e : Probability of random agreement


P_e : Probability of random agreement
likes 70% of movies
likes 50% of movies

P_e : Probability of random agreement

likes 70% of movies

likes 50% of movies

P_e : Probability of random agreement

$P(good_{cat})$ =

likes 70% of movies

$P(good_{penguin})$ =

likes 50% of movies

P_e : Probability of random agreement

$P(good_{cat}) = 0.7$

likes 70% of movies

$P(good_{penguin}) =$

likes 50% of movies

P_e : Probability of random agreement

$P(good_{cat}) = 0.7$

likes 70% of movies

$P(good_{penguin} \cap good_{cat})$

= ????

$P(good_{penguin}) = 0.5$

likes 50% of movies

P_e : Probability of random agreement

$P(good_{cat}) = 0.7$

likes 70% of movies

$P(good_{penguin} \cap good_{cat})$

$= 0.7 \times 0.5$

$= 0.35$

$P(good_{penguin}) = 0.5$

likes 50% of movies

Cohen's Kappa

this is just % agreement

P_0 = relative observed agreement

P_e = probability of random agreement

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

this what we just did

P_e : Probability of random agreement (pick up line)

The proportion of times you'd expect annotators to agree if they were each randomly assigning categories, but using their own observed **base rates.**

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

P_e : Probability of random agreement (pick up line)

The proportion of times you'd expect annotators to agree if they were each randomly assigning categories, but using their own observed **base rates.**

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- What's the **base rate** P(pick up line) for Person 1?
- What's P(pick up line) for person 2?

P_e : Probability of random agreement (pick up line)

The proportion of times you'd expect annotators to agree if they were each randomly assigning categories, but using their own observed **base rates.**

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- What's the **base rate** P(pick up line) for Person 1?
  - = 30 / 50
  - = 3 / 5
  - = 0.6
- What's P(pick up line) for person 2?
  - = 25 / 50
  - = 1 / 2
  - 0.5

P_e : Probability of random agreement (pick up line)

The proportion of times you'd expect annotators to agree if they were each randomly assigning categories, but using their own observed **base rates.**

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- What's the **base rate** P(pick up line) for Person 1?
  - = 30 / 50
  - = 3 / 5
  - = 0.6
- What's P(pick up line) for person 2?
  - = 25 / 50
  - = 1 / 2
  - 0.5
- What's the base rate of agreement on whether or not something **is** a pickup line?

P_e : Probability of random agreement (pick up line)

The proportion of times you'd expect annotators to agree if they were each randomly assigning categories, but using their own observed **base rates.**

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- What's the **base rate** P(pick up line) for Person 1?
  - = 30 / 50
  - = 3 / 5
  - = 0.6
- What's P(pick up line) for person 2?
  - = 25 / 50
  - = 1 / 2
  - 0.5
- What's the base rate of agreement on whether or not something **is** a pickup line?
  - = 0.6 x 0.5
  - = 0.30

P_0 : Not done yet

We now know that 30% of the time, Person 1 and Person 2 will just happen to agree that it's a pickup line.

This is just one kind of agreement: both agree that "pickup line"

We need to calculate this same chance probability for the other kind of agreement. The case where both agree that "trash talk"

P_e : Probability of random agreement (trash talk)

The proportion of times you'd expect annotators to agree if they were each randomly assigning categories, but using their own observed **base rates.**

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- What's the **base rate** P(trash talk) for Person 1?
- What's P(trash talk) for person 2?

P_e : Probability of random agreement (trash talk)

The proportion of times you'd expect annotators to agree if they were each randomly assigning categories, but using their own observed **base rates.**

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- What's the **base rate** P(trash talk) for Person 1?
  - = 20 / 50
  - = 2 / 5
  - = 0.4
- What's P(trash talk) for person 2?
  - = 25 / 50
  - = 1 / 2
  - 0.5

P_e : Probability of random agreement (trash talk)

The proportion of times you'd expect annotators to agree if they were each randomly assigning categories, but using their own observed **base rates.**

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- What's the **base rate** P(trash talk) for Person 1?
  - = 20 / 50
  - = 2 / 5
  - = 0.4
- What's P(trash talk) for person 2?
  - = 25 / 50
  - = 1 / 2
  - 0.5
- What's the base rate of agreement on whether or not something **is not** a pickup line (because it's trash talk)?

P_e : Probability of random agreement (trash talk)

The proportion of times you'd expect annotators to agree if they were each randomly assigning categories, but using their own observed **base rates.**

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

- What's the **base rate** P(pick up line) for Person 1?
  - = 30 / 50
  - = 3 / 5
  - = 0.6
- What's P(pick up line) for person 2?
  - = 25 / 50
  - = 1 / 2
  - 0.5
- What's the base rate of agreement on whether or not something **is not** a pickup line (because it's trash talk)?
  - = 0.4 x 0.5
  - **= 0.20**

P_e : Probability of random agreement

So we predict that **20%** percent of the cases A and B will just *happen* to agree that it **is trash talk,** and **30%** of the time A and B will just *happen* to agree that it is **not trash talk** (because it's a pick up line).

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

What percentage of the time can we expect A and B to agree?

P_e : Probability of random agreement

So we predict that **20%** percent of the cases A and B will just *happen* to agree that it **is trash talk,** and **30%** of the time A and B will just *happen* to agree that it is **not trash talk** (because it's a pick up line).

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

What percentage of the total cases should we expect A and B to agree on?

20% + 30% = 50% of the time!

$P_e = 0.5$

Cohen's Kappa

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

P_0 = relative observed agreement

P_e = probability of random agreement

$P_0 = 0.7$

$P_e = 0.5$ 

So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

A \ B	Yes	No
Yes	a	b
No	c	d

For reference

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

e.g.

A \ B	Yes	No
Yes	20	5
No	10	15

So the expected probability that both would say yes at random is:

$$p_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = 0.5 \times 0.6 = 0.3$$

Similarly:

$$p_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d} = 0.5 \times 0.4 = 0.2$$

Overall random agreement probability is the probability that they agreed on either Yes or No, i.e.:

$$p_e = p_{\text{Yes}} + p_{\text{No}} = 0.3 + 0.2 = 0.5$$

So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

P_e : Probability of random agreement

percent agreement (same as P_0) = 0.7

Cohen's K = 0.4

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

What is Cohen's Kappas doing?

percent agreement (same as P_0) = 0.7

Cohen's K = 0.4

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

Why are they different?

What does Cohen's Kappa take into account that percent agreement does not?

What is Cohen's Kappas doing?

percent agreement (same as P_0) = 0.7

Cohen's K = 0.4

		Person 1: ____Sooji____		
		label a: _pick up line__	label b: ___trash talk__	total:
Person 2: __Gabriella__	label a: _pick up line__	20	5	25
	label b: ___trash talk__	10	15	25
	total:	30	20	total N = 50

Why are they different?

What does Cohen's Kappa take into account that percent agreement does not?

Kappa factors out the probability of agreement due to chance