MMA Inter-rater Reliability Data Analysis

Caleb Skinner; Joshua Patrick, PhD; Connor Bryson; Rodney X. Sturdivant, PhD

Contents

Introduction **2**

Learning Objectives 2

Inter-rater Reliability Overview **2**

MMA Overview **3**

Data **3**

Measuring Inter-rater Reliability **4**

Percent Agreement 4

Cohen's Kappa 11

Weighted Kappa 17

Fleiss' Kappa 25

Conclusion **29**

Introduction

What do mixed martial arts, figure skating, medical diagnoses, and essay grading all have in common? Not much on the surface; except they all use human judges to evaluate or measure an event. Each of these fields- and many more- rely heavily on human judgment for their results. Sometimes, a different judge could lead to an entirely new conclusion. How do we know when to trust our human measurement systems? How much can we trust them?

In mixed martial arts (MMA) when a fight is not decided by a knockout the winner is determined by a panel of judges. Sometimes the decision is unanimous - all the judges agree about which fighter won the fight. However, sometimes the judges do not agree and a "split decision" is produced. In some cases, the decision is controversial especially when the judges disagree.

If fans and fighters begin to question decisions and fairness, the sport is hurt. The UFC and other governing bodies desire judging that is consistent and reliable to produce trust in the scoring of fights.

In this module, you will explore statistics used to measure judges' consistency and reliability and apply them to judging of MMA fights. The analysis could inform the fight commissions about issues with specific judges and fairness in scoring fights more broadly.

Learning Objectives

By the end of this module you will be able to:

- understand the basic concepts of inter-rater reliability.
- understand the purposes of inter-rater reliability
- understand and interpret measures of reliability:
  - Percent agreement
  - Cohen's Kappa
  - Weighted Kappa
  - Fleiss' Kappa
- apply reliability measures to judging of MMA fight data

Inter-rater Reliability Overview

Inter-rater reliability measures the consistency of two or more individuals rating something. Simply put, it measures the extent that these judges agree in their ratings. Imagine several teachers grading a single essay. In a perfect world, all of the teachers will award the essay the same score. Inter-rater-reliability measures the consistency of these "raters" when rating the same thing.

Several different judges can rate the same event and we would expect similar results. If the judges' scores are not similar, then the system may be flawed. Ideally, the particular judge(s) of an event do not have a significant impact on the result. If the measurements are not reliable, different set of judges could yield an entirely different conclusion.

This process also gives us information on the judges themselves. If there is a high consistency between the judges- then we can say the judges are **interchangeable**. In other words, substituting one judge for another would lead to little change in the results. Some individuals may tend to differ from the other judges more often and inter-rater reliability can also help us identify those raters.

However, inter-rater-reliability can help us to test the reliability of a measurement system or interchangeability of judges, but it cannot speak to its **validity** or accuracy. While unlikely, it is conceivable that results

from a group of judges could be consistent and reproducible, yet inaccurate. Testing the accuracy of the data is beyond the scope of inter-rater reliability.

Inter-rater reliability should not be confused with *intra*-rater reliability. *Inter*-rater reliability measures the variation between multiple judges evaluating one event. *Intra*-rater reliability measures the variation of one judge evaluating multiple trials of one event. They are both important, and they often use similar methods, but we will be focusing on *inter*-rater reliability today.

MMA Overview

As aforementioned, inter-rater reliability can be implemented in judge-based sports. One of the most well-known of these sports is Mixed Martial Arts or MMA. In this module, we'll walk you through some inter-rater reliability methods using results from MMA competitions. MMA encompasses all sorts of fighting methods between two fighters. Some popular forms are boxing, kickboxing, muay thai, jiu-jitsu, and wrestling. These forms have different fighting styles and rules, but they generally have the same outcome structure.

In simple terms, the fights always end in one of two ways: first, one fighter is unable to continue, or second, the time runs out. It is easy to determine a winner in the first scenario. The second is more difficult. Most fights are three rounds, but championship fights and some main event fights are five rounds. After the final bell, the outcome rests in the hands of three judges to score the match and determine a winner. MMA terms this act a "decision".

Organizations like the UFC employ these judges, but they cannot select them. This responsibility resides with Athletic Commissions - like the Nevada State Athletic Commission or New York State Athletic Commission - to delegate judges to the events. To prevent corruption, judges are separate from the organization that holds the events. Judges are also kept independent from each other. They sit in different booths in the arena and score the fights without consulting one another.

Data

Our data includes 5000 fights that end in a decision. All 5000 of these fights reached the final bell without a natural victor. The three judges scored the fight and determined a winner.

The data spans from 2001 to 2021, including fights from all over the world and many organizations like UFC, Bellator, Invicta, Strikeforce, and World Extreme Cagefighting. Here is a list of the variables in our data set, and a data dictionary to help you understand them. Fighter 1 and fighter 2 were randomly assigned these labels. There is no meaning behind this classification.

Fight rules change over time which could impact judging. We have not attempted to address this issue in our data collection or analysis in introducing inter-rater reliability. Further work to take changes into account would be interesting.

variables	explanation	example
date	date of fight in year - month - day format	2012-02-26, 2019-10-06, etc.
rounds	number of rounds in the fight	3 or 5
fighter1	last name of the first fighter	Soto, VanZandt, etc.
fighter2	last name of the second fighter	Rivera, Delboni, etc.
winner	winner of the match	Rivera, VanZandt, Draw, etc.

variables	explanation	example
result_type	decision scoring terminology	Unanimous (all agree), Split (at least one judge votes for each fighter), Majority (two judges votes for a fighter, one judge votes for a draw)
judgen	last name of the judge n	Chatfield, Collett, etc.
judgen_score1	judge n's score for fighter 1	30, 28, etc.
judgen_score2	judge n's score for fighter 2	27, 29, etc.
judgen_margin	difference in judge n's score for fighters 1 and 2	30-27 = 3, 28-29 = -1, etc.
judgen_out	outcome of judge n's decision	fighter1, draw, fighter2

Here is a slice of the data (the first 9 columns with judge 1 scores).

date	rounds	fighter1	fighter2	winner	result_type	judge1	judge1_score1	judge1_score2
2021-12-18	5	Torres	Parnasse	Parnasse	Unanimous	Motylewski	45	50
2021-12-18	3	Eskiev	Stasiak	Eskiev	Unanimous	Motylewski	29	27
2021-12-18	3	Bekus	Sormova	Bekus	Unanimous	Motylewski	29	28
2021-12-18	3	Rewera	Erzanukaev	Erzanukaev	Unanimous	Motylewski	26	30
2021-12-18	3	Thompson	Muhammad	Muhammad	Unanimous	Bell	25	30
2021-12-18	3	Lemos	Hill	Lemos	Split	Bell	28	29
2021-12-18	3	Ewell	Jourdain	Jourdain	Unanimous	D'Amato	27	30
2021-12-18	3	Niedzwiedz	Broz	draw	Majority	Motylewski	28	28
2021-12-11	3	Figlak	Kauppinen	Figlak	Unanimous	Brown	30	27

In our data, there are three or five rounds of competition. Each of the three judges score both fighters on each round. A typical victory in a round gives 10 points for the victor and 9 points for his opponent. A large victory awards 10 points and 8 points, and an overwhelming victory awards 10 points and 7 points, respectively. Note that scores of 6 and below are not possible.

After totaling up the points, each judge arrives at his or her outcome. A victory for the first fighter, victory for the second fighter, or a draw. However, as you can guess, they often disagree.

Important note: *In this module we will use the total score for the fight to judge rater reliability. Judges may "agree" in their total score but not agree on the round to round scoring. For example, Judge 1 could score rounds 10-9, 10-9, 9-10 and Judge 2 9-10, 10-9, 10-9 leading to the same overall score of 29-28. However, the judges disagreed on two of the three rounds of the fight. Thus, the analysis here is likely to produce higher reliability scores than if we used the individual round scores.*

Measuring Inter-rater Reliability

Percent Agreement

We'll begin by evaluating the inter-rater-reliability between two judges. One simple way is to calculate the percentage of fights in which they agree on the outcome. This is appropriately termed **percent agreement**.

This reduces each fight into a simple outcome: agree or disagree, so we lose any information on the scores or margin of the fight (we will return to the original scores later.) Another limitation of percent agreement is that it ignores the possibility of judges arriving at agreement through chance. Still, it provides a foundation for the widely used **Cohen's kappa**. There will be more on this later, but for now, let's calculate the percent agreement for some judges.

Below, we see a table of the ten most frequently-appearing judges. The table includes the number of fights that they judged. D'Amato, Lee, and Cleary are the most experienced judges in our data set.

judge	fights
D'Amato	846
Lee	534
Cleary	532
Cartlidge	413
Weeks	384
Bell	383
Colón	372
Crosby	368
Rosales	321
Lethaby	305

Example: D'Amato and Lee

Let's take the top two judges: D'Amato and Lee, and compare their rulings on fights where they both judged the same fight. We'll look specifically at the outcome categorical variable: judge_out.

Here is a **contingency table** that summarizes all the fights that D'Amato and Lee judged together. Contingency tables are used to assess the association between two paired categorical variables. They tabulate the distribution of each variable and compare their results.

Table 4: D'Amato and Lee's Decisions

		Lee		
D'Amato	**fighter1**	**draw**	**fighter2**	**total**
fighter1	62	0	12	**74**
draw	4	2	0	**6**
fighter2	10	0	52	**62**
total	**76**	**2**	**64**	**142**

This table helps us to organize and compare D'Amato's and Lee's ratings. Each row consists of D'Amato's votes and each column consists of Lee's corresponding votes. Thus, the table forms a downward sloping diagonal where the judges agree. These are called **concordant responses**.

They both selected fighter 1 as the victor 62 times, a draw twice, and fighter 2 as victor 52 times. However, they selected opposing fighters 22 times (10 + 12), and D'Amato voted for a draw four times where Lee disagreed.

Our total concordant values is $62 + 2 + 52 = 116$. We can calculate the simple percent agreement by adding up all the concordant values, and dividing it by the total number of results. Another word for this is the **proportion of observed agreement** (p_o). This term will be important later.

Table 5: D'Amato and Lee - Simple Agreement

Agree	Disagree	Percent Agreement
116	26	81.69%

D'Amato and Lee's simple agreement (proportion of observed agreement) is only 81.69%. This means they disagreed on almost 1/5 of their rulings.

Exercise 1: Percent Agreement for D'Amato and Cleary

Let's compare D'Amato and Lee's consistency with that of D'Amato and Cleary. Below is the contingency table of D'Amato and Cleary.

Table 6: D'Amato and Cleary's Decisions

	Cleary			
D'Amato	**fighter1**	**draw**	**fighter2**	**total**
fighter1	65	0	4	**69**
draw	1	3	1	**5**
fighter2	7	0	71	**78**
total	**73**	**3**	**76**	**152**

1.1. Can you identify D'Amato and Cleary's concordant responses?

SOLUTION: 139

1.2 Use the concordant responses to calculate D'Amato and Cleary's percent agreement yourself. Remember, the formula for percent agreement (p_o) = concordant values/total.

SOLUTION: 91.4%

1.3. We've left the table empty for you. Fill it in with the appropriate values.

SOLUTION: 139/13/91.4%

Table 7: D'Amato and Cleary - Simple Agreement

Agree	Disagree	Percent Agreement

1.4 How do the results compare with D'Amato and Lee?

SOLUTION: D'Amato and Lee tend to agree less often

Exercise 2: Cartlidge and Lethaby

We next consider an example using the top judge combination Cartlidge and Lethaby.

Table 8: Cartlidge and Lethaby's Decisions

	Lethaby			
Cartlidge	**fighter1**	**draw**	**fighter2**	**total**
fighter1	88	0	16	**104**
draw	2	1	3	**6**
fighter2	7	0	96	**103**
total	**97**	**1**	**115**	**213**

2.1. Assess the table. How well do the judges tend to agree?

SOLUTION: Reasonable agreement (less than 30 times disagree)

2.2. Once again, we've left an empty percent agreement table for you. Fill in the cells with the appropriate values and calculate the percent agreement.

SOLUTION: 186/28/86.8%

Table 9: Cartlidge and Lethaby - Simple Agreement

Agree	Disagree	Percent Agreement

2.3. How do Cartlidge and Lethaby's results compare to the other judges we assessed?

SOLUTION: They have slightly better agreement than D'Amato and Lee but not as strong as D'Amato and Cleary

Exercise 3 - Advanced (Optional): Other Judges

Use R code (or the package you use for data analysis) to select any two judges to compare from the list below.

judge1	judge2	fights_judged
Cartlidge	Lethaby	213
Cleary	D'Amato	152
D'Amato	Lee	142
Collett	Lethaby	116
Cartlidge	Collett	113
Cleary	Lee	98
Colón	Tirelli	90
Bell	Mccarthy	78

judge1	judge2	fights_judged
D'Amato	Kamijo	75
D'Amato	Weeks	74
Crosby	D'Amato	70
Bell	D'Amato	69
Bell	Cleary	67
Kamijo	Weeks	64
Colón	D'Amato	63

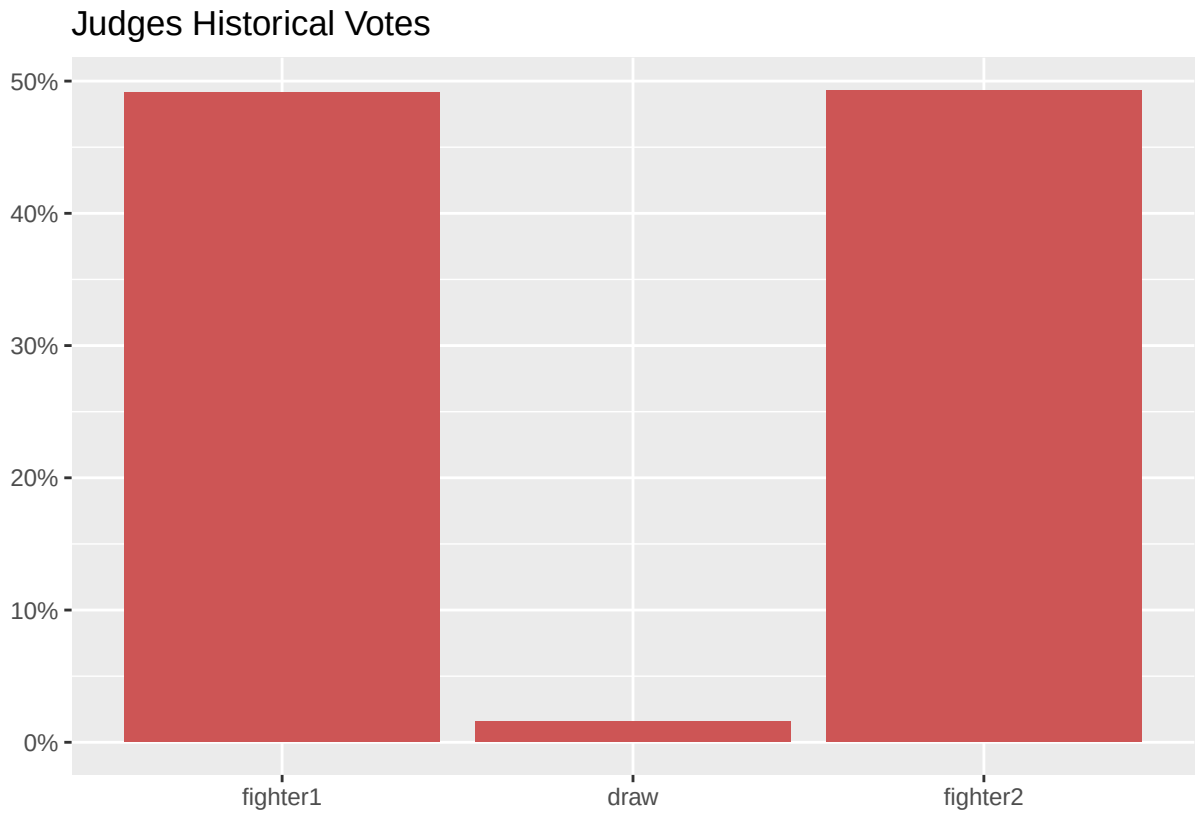
Limitations of Percent Agreement

Percent agreement is helpful, because it gives us a general understanding of the judges' reliability, but it is limited. In particular, it cannot account for the judges' arriving at similar conclusions via chance.

So, how likely is it for judges to arrive at similar conclusions via chance even if they do not necessarily judge consistently?

Let's look at a simple simulation. Here, we have two hypothetical judges - we'll call them Jimmy and Mateo - rating 1000 fights. Except, instead of watching and analyzing the fights before carefully determining a winner, both Jimmy and Mateo slept through all 1000 fights. Luckily for them, they remembered the historical voting trends of MMA judges. Both of them, independently, decided to randomly select a winner for each of the 1000 fights in a way that was consistent with the likelihoods of the historical rulings.

The historical rulings are below. In the 1000 fights, they selected fighter 1 and fighter 2 about 49% of the time and a draw about 2% of the time.



They made these ratings without consulting each other or watching the fights. Below are the first 15 observations of the data set we created.

fight	Jimmy	Mateo
1	fighter1	fighter2
2	fighter2	fighter2
3	fighter2	fighter2
4	fighter2	fighter1
5	fighter1	fighter2
6	fighter1	fighter1
7	fighter1	fighter1
8	fighter1	fighter1
9	fighter1	fighter1
10	fighter2	fighter2
11	fighter2	fighter1
12	fighter2	fighter2
13	fighter1	fighter2
14	fighter2	fighter2
15	fighter2	fighter1

And here is their contingency table.

Table 12: Jimmy and Mateo's Decisions

	Mateo			
Jimmy	**fighter1**	**draw**	**fighter2**	**total**
fighter1	241	12	248	**501**
draw	9	0	13	**22**
fighter2	224	6	247	**477**
total	**474**	**18**	**508**	**1,000**

Their simple agreement numbers come out like this:

Agree	Disagree	Percent Agreement
488	512	48.80%

Woah! Jimmy and Mateo agreed *48.80%* of the time. This certainly is not a good rate of agreement, but it does suggest caution in interpreting percent agreement rates of our real judges D'Amato, Lee, and Cleary. Their percent agreements fell in the 80-90% range, but we can get over half that agreement with just random chance.

We next consider metrics that account for the fact that some agreement is likely just due to chance.

Cohen's Kappa

Cohen's kappa is a second, more rigorous method, that assesses the agreement between two judges. Like percent agreement, it measures the reproducibility of repeated assessments of the same event. It was developed by Jacob Cohen in the 1960s as an alternative agreement method that accounts for the possibility of chance agreement.

Cohen's kappa makes a few assumptions about the data:

1. The same two individuals must rate each event.
2. The principle of **independence**. The judges rate the same events without consultation or communication. This means the judges' results are **paired**.
3. The judgments are made between the same defined categories. In our context, the judges categorize the fight result as win/lose/draw for each fighter.

All three of these assumptions are met by our data. We will filter our data to ensure the same two judges score each event. Judges in MMA fights are kept in separate areas around the fight. All our judges vote for fighter 1, fighter 2, or a draw.

Like percent agreement, Cohen's kappa works with any categorical variable.

Cohen's kappa isolates the judges' real agreement from their chance agreement. It produces a correlation coefficient kappa (κ) that assesses the agreement between the two judges and ranges from -1 to 1.

- At κ = -1, the two judges produced exactly opposite assessments of the event.
- At $\kappa = 0$, the agreement between the two judges is tantamount to an agreement entirely produced by chance.
- At $\kappa = 1$, the two judges have perfect agreement. Their assessments of the events are identical.

Example: D'Amato and Lee

As we walk through the methodology of Cohen's kappa, let's revisit our example of Lee and D'Amato.

Again, we begin with a contingency table.

Table 14: D'Amato and Lee's Decisions

		Lee		
D'Amato	**fighter1**	**draw**	**fighter2**	**total**
fighter1	62	0	12	**74**
draw	4	2	0	**6**
fighter2	10	0	52	**62**
total	**76**	**2**	**64**	**142**

Earlier, we found the proportion of observed agreement for this table is 81.69%. If we're going to account for chance, we need to estimate what the agreement rate would be if the results were completely randomized.

We can estimate these random results by producing theoretical estimates. This is called the **expected value**. We calculate the expected value of each cell by multiplying together three values. The first judge's probability of producing a result, the second judge's probability of producing the corresponding result independent of the first judge, and the total number of fights.

For example, to find the expected value in the draw-draw concordant cell. We can multiply D'Amato's draw rate of $\frac{6}{142}$ by Lee's draw rate of $\frac{2}{142}$ and by the total number of fights: 142. We end up with 0.085.

In other words, If D'Amato and Lee were to judge a new set of 142 fights and randomly pick their results from a hat containing their historical results together, we'd expect them to both pick a draw 0.085 times.

We created a table full of the expected values.

Table 15: Expected Values of D'Amato and Lee

		Lee		
D'Amato	**fighter1**	**draw**	**fighter2**	**total**
fighter1	39.61	1.04	33.35	**74**
draw	3.21	0.08	2.70	**6**
fighter2	33.18	0.87	27.94	**62**
total	**76.00**	**2.00**	**64.00**	**142**

With the table, we sum up the three concordant cells: 67.63, and divide by the total number of fights: 142. This gives us the **proportion of expected agreement** (p_e). A value of 47.63%.

Now, with both the proportion of observed agreement and the proportion of expected agreement, we can calculate kappa using the formula:

- $kappa(\kappa) = \frac{p_o - p_e}{1 - p_e}$

When we plug in those values and solve for kappa, we find that the kappa between D'Amato and Lee is 0.65.

We're past all the calculations and math, but what does our kappa mean?

Interpreting Kappa

The kappa value represents the percentage of the two judges' results that agree with one another over and above what we would expect from chance. Conversely, the complement of kappa represents the percentage of the two judge's results that result from chance or straight-up disagreement.

Thus, the magnitude of the agreement is important. A higher kappa is always better, because it suggests a higher reproducibility in the measurement system. Unlike some statistical tests, the kappa statistic is not evaluated by passing a threshold. Instead, the exact assessment of a kappa often depends on a myriad of factors.

This contextual nature of kappa makes interpretation difficult. There is a lot of disagreement over the interpretations for different kappa values, and the guidelines typically vary with the field of study. For example, health related studies demand a stronger reliability than fields that have less widespread influence over the population's well-being like, say, judging MMA fights.

Below is one evaluation method, that has been generally agreed upon by several prominent statisticians:

- $\kappa > 0.75$ Excellent reproducibility
- $0.4 \leq \kappa \leq 0.75$ Good reproducibility
- $0 \leq \kappa < 0.4$ Marginal reproducibility

Note that a negative kappa value is possible, and would represent agreement that is worse than that expected by chance. Clearly such a value would indicate very poor reproducibility.

Applying this method to our judges, D'Amato and Lee's kappa of 0.65 indidcates the judges have "good" reproducibility in their ratings. They have decent consistency and interchangeability. We should be careful, because an estimated 35% of their relationship is comprised of chance agreement or disagreement. However, we cannot speak to D'Amato and Lee's accuracy or validity in their ratings. We cannot assess if their judgments were correct.

Advanced (optional): Confidence intervals and tests for Kappa

We can produce a confidence interval and hypothesis test for our kappa.

With a large enough sample size, kappa is normally distributed with a standard error (se).

- $se(\kappa) = \sqrt{\frac{1}{n(1-p_e)^n} * [p_e + p_e^2 - \sum_{i=1}^{c} (a_i b_i (a_i + b_i))]}$

Using this standard error, we can calculate the 95% confidence interval by:

- $\kappa = \pm 1.96 * se(\kappa)$

A 95% confidence interval produces an estimated range for the true value of kappa. We can say with 95% confidence that the interval includes the true kappa. Like all confidence intervals, a larger sample size reduces this interval.

The confidence interval is important, because it helps us to see how much we can trust our kappa. A high kappa statistic that has a large confidence interval is far from ideal.

Our 95% confidence interval for the kappa of D'Amato and Lee is 0.529 to 0.772. Thus, we can say with 95% confidence that the interval (0.529, 0.772) includes the true value of kappa. In other words, we would not be surprised if the true kappa is as low as 0.529.

We can also create a hypothesis test for our kappa. We're looking to test that there is at least some non-random association between the judges.

Our kappa test has a null and alternative hypothesis of:

- $H_o : \kappa = 0$
- $H_a : \kappa > 0$

We will hold to our null hypothesis unless we have significant evidence to reject it. This evidence is held in a p-value. If our p-value is less than our α of 0.05, then we have sufficient evidence to reject our null hypothesis and agree with our alternative. Moreover, if our confidence interval does not include 0 within its range, then we can reject the null hypothesis without checking for the p-value.

The hypothesis test can be misleading, however, because a small kappa value can reject the null hypothesis despite indicating only poor agreement. For this reason, confidence intervals are preferable.

Our p-value for D'Amato and Lee is 2.22×10^{-16}. We can thoroughly reject the null hypothesis that there is no association in the decisions of D'Amato and Lee.

Exercise 4: D'Amato and Cleary

Now that we've analyzed D'Amato and Lee. We'd like to give you the opportunity to describe the agreement between D'Amato and Cleary. We'll produce the results and you produce the analysis.

Table 16: D'Amato and Cleary's Decisions

		Cleary		
D'Amato	**fighter1**	**draw**	**fighter2**	**total**
fighter1	65	0	4	**69**
draw	1	3	1	**5**
fighter2	7	0	71	**78**
total	**73**	**3**	**76**	**152**

Table 17: Expected Values of D'Amato and Cleary

		Cleary		
D'Amato	**fighter1**	**draw**	**fighter2**	**total**
fighter1	33.14	1.36	34.5	**69**
draw	2.40	0.10	2.5	**5**
fighter2	37.46	1.54	39.0	**78**
total	**73.00**	**3.00**	**76.0**	**152**

4.1. Recall our percent agreement assessment from earlier. Do D'Amato and Cleary tend to agree?

SOLUTION: Very good agreement (less than 15 times disagree)

4.2. Compare the two tables. Do the expected values for the cells surprise you?

SOLUTION: No. Expected values are based on random selections so we would expect about 50 percent agreement

4.3. After some calculations, we find that D'Amato and Cleary's kappa is 0.84. Using the guidelines demonstrated above, interpret the value.

SOLUTION: This value suggests excellent reproducibility

4.4. (Optional - Advanced) We can run the confidence interval and hypothesis test through our program:

```
##
##  Estimate Cohen's kappa statistics and test the null hypothesis that the
##  extent of agreement is same as random (kappa=0)
##
## data:  .
## Z = 10.844, p-value < 2.2e-16
## 95 percent confidence interval:
##  0.7522942 0.9217408
## sample estimates:
## [1] 0.8370175
```

Analyze the results. Produce explanations of the confidence interval and hypothesis test, and provide your own assessment of the association between D'Amato and Cleary. Try to use the wording and phrases that we explained earlier.

SOLUTION: The interval is entirely in the excellent category so we can reject a null hypothesis of less than excellent reproducability

Exercise 5: Cleary and Lee

Next, we have selected to analyze Cleary and Lee.*Note: the lower the sample size, the wider our confidence interval and the less we can trust our kappa value.*

5.1 How does the sample size between these two judges likely impact the kappa estimate?

SOLUTION: The sample size is smaller than previous examples so we will have LESS confidence in the kappa estimate

Table 18: Cleary and Lee's Decisions

		Lee		
Cleary	**fighter1**	**draw**	**fighter2**	**total**
fighter1	44	1	11	**56**
draw	2	0	0	**2**
fighter2	7	0	33	**40**
total	**53**	**1**	**44**	**98**

Table 19: Expected Values of Cleary and Lee

		Lee		
Cleary	**fighter1**	**draw**	**fighter2**	**total**
fighter1	30.29	0.57	25.14	**56**
draw	1.08	0.02	0.90	**2**
fighter2	21.63	0.41	17.96	**40**
total	**53.00**	**1.00**	**44.00**	**98**

5.2. Assess the two tables. Do the judges appear to agree?

SOLUTION: They do disagree a fair amount relative to the sample size, but still near 80 percent

5.3. Now, compare the two tables. Do the expected values for the cells surprise you? How similar are they to the observed values?

SOLUTION: Expected agreement is roughly 50 percent so not surprising.

5.4. After solving for kappa, our value is 0.21. Using the guidelines demonstrated previously, interpret the value. How does it compare to previous judge-pairings kappas?

SOLUTION: This value suggests marginal reproducibility, much worse than previous examples

```
##
##  Estimate Cohen's kappa statistics and test the null hypothesis that the
##  extent of agreement is same as random (kappa=0)
##
## data:  .
## Z = 4.1376, p-value = 1.755e-05
## 95 percent confidence interval:
##  0.08973446 0.32838582
## sample estimates:
## [1] 0.2090601
```

5.5. (Optional advanced) Interpret the confidence interval and p-value. What does they mean for the relationship between the two judges?

SOLUTION: We have little evidence that the true value might be better than marginal agreement.

Weighted Kappa

Great. We've analyzed our data and produced a kappa value that assesses the true agreement between judges while accounting for random chance.

Still, we are losing some information. Our judges supply score cards with point values for each fighter. They don't just assign a winner. When we reduce each judge's ruling to win, lose, or draw, we miss out on the degree of these victories. Instead of looking at the outcome variable, let's analyze the margin variable.

In this case, the margin variable is an **ordinal** variable. Ordinal variables are a type of categorical variable that has a similar function to nominal variables, except that there is a clear ordering in the results. Height, for example, can be divided into ordinal categories like "very tall", "tall", "normal", "short", and "very short".

This clear ordering of the categories allows for **partial agreement**. Partial agreement affords some credit to close responses. The judge's responses may not be identical, but they could be close. Short is a lot closer to very short than very tall. Partial agreement takes this into account.

Weighted kappa is a variant of Cohen's kappa (also created by Jacob Cohen) that permits this partial agreement between responses. Like Cohen's kappa, it accounts for any chance agreement, but it also takes into account the proximity of the judges' results. A large disparity in the two judge's margin will lower the agreement much more than smaller disparities. The unweighted Cohen's kappa, however, treats all disparities equally.

The weighted kappa makes assumptions that are similar to Cohen's kappa about the data:

1. The same two individuals must rate each event.
2. The principle of **independence**. The judges rate the same events without consultation or communication. This means the judges' results are **paired**.
3. The judgments are made between the same ordinal categories.

Exercise 6: D'Amato and Lee

Weighted kappa begins like the Cohen's kappa with a contingency table. To simplify the analysis for this example, we only kept the fights that went three rounds.

Table 20: D'Amato and Lee's Decision Margins

						Lee						
D'Amato	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
-5	0	1	0	0	0	0	0	0	0	0	0	**1**
-4	0	4	0	0	0	0	0	0	0	0	0	**4**
-3	1	0	13	0	4	0	0	0	0	0	0	**18**
-2	0	0	1	2	1	0	0	0	0	0	0	**4**
-1	0	0	2	1	14	0	9	0	1	0	0	**27**
0	0	0	0	0	0	2	3	0	0	0	0	**5**
1	0	0	1	0	7	0	18	1	5	0	0	**32**
2	0	0	0	0	0	0	1	2	0	0	0	**3**
3	0	0	1	0	0	0	6	0	11	2	0	**20**

Table 20: D'Amato and Lee's Decision Margins

						Lee					
D'Amato	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
4	0	0	0	0	0	0	0	0	0	3	0	**3**
5	0	0	0	0	0	0	0	0	1	1	0	**2**
Total	**1**	**5**	**18**	**3**	**26**	**2**	**37**	**3**	**18**	**6**	**0**	**119**

6.1. Take a look at the contingency table. Trace your eyes along the diagonal of concordant values. How often do the judges completely agree?

SOLUTION: 69 times, 58 percent

6.2. In the first column, there is a single observation (a 1 in row three). What does this value represent?

SOLUTION: Agreement about the fight outcome (winner) but Lee had a wider margin in scores

6.3. What are the most common frequencies? Why?

SOLUTION: 0. There are many possible score combinations for the size of the sample.

Table 21: Expected Values of D'Amato and Lee

						Lee					
D'Amato	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
-5	0.0	0.0	0.2	0.0	0.2	0.0	0.3	0.0	0.2	0.1	0	**1**
-4	0.0	0.2	0.6	0.1	0.9	0.1	1.2	0.1	0.6	0.2	0	**4**
-3	0.2	0.8	2.7	0.5	3.9	0.3	5.6	0.5	2.7	0.9	0	**18**
-2	0.0	0.2	0.6	0.1	0.9	0.1	1.2	0.1	0.6	0.2	0	**4**
-1	0.2	1.1	4.1	0.7	5.9	0.5	8.4	0.7	4.1	1.4	0	**27**
0	0.0	0.2	0.8	0.1	1.1	0.1	1.6	0.1	0.8	0.3	0	**5**
1	0.3	1.3	4.8	0.8	7.0	0.5	9.9	0.8	4.8	1.6	0	**32**
2	0.0	0.1	0.5	0.1	0.7	0.1	0.9	0.1	0.5	0.2	0	**3**
3	0.2	0.8	3.0	0.5	4.4	0.3	6.2	0.5	3.0	1.0	0	**20**
4	0.0	0.1	0.5	0.1	0.7	0.1	0.9	0.1	0.5	0.2	0	**3**
5	0.0	0.1	0.3	0.1	0.4	0.0	0.6	0.1	0.3	0.1	0	**2**
Total	**1.0**	**5.0**	**18.0**	**3.0**	**26.0**	**2.0**	**37.0**	**3.0**	**18.0**	**6.0**	**0**	**119**

6.4. Now look at the expected values. What values are the largest? Is this surprising?

SOLUTION: All less than 10. There are many possible score combinations for the size of the sample.

6.5. Does the observed agreement surpass the expected agreement? By how much?

SOLUTION: Yes. 58 observed vs 18 percent expected.

Weights

Like Cohen's kappa, the weighted kappa calculates the proportion of observed agreement and the proportion of expected agreement by using the concordant values along the diagonal of our contingency tables. However, the calculation of these two agreements becomes more complex, because we can allow for partial agreement for close matches. The formulas are the same as Cohen's kappa, except for the addition of the weights:

- $p_o = \sum_i \sum_j W_{ij} P_{ij}$
- $p_e = \sum_i \sum_j W_{i+} P_{+j}$

where W is the weight for each cell and P is the proportion of each cells frequency.

The weights W are proportions between 0 and 1 that reflect the level of agreement. All concordant values have complete agreement, so their weight is 1. Values to the left and right of the diagonal have proprtions slightly less than 1 and so on. In the standard unweighted Cohen's kappa, all the diagonal values have weights of 1 and the non-diagonal values have weights of 0.

There are many different ways to calculate the weights and selecting them generally depends on the size of the table and the distribution of the variables. Two common methods are linear and quadratic weighting.

Linear weights, formally known as Cicchetti-Allison weights, create equal distance between the weights. A cell's weight is directly proportional to its distance from the concordant value.

The formula for the linear weights are:

- $W_{ij} = 1 - (|i - j|)/(R - 1)$

R is the total number of categories and |i - j| is the distance between the two cells.

Let's calculate the weights of the first few cells using the the formula. We'll begin with the (-5, -5) cell and move right on the table.

- $W_{-5,-5} = 1 - (|0|/(11 - 1)) = 1$
- $W_{-5,-4} = 1 - (|1|/(11 - 1)) = .9$
- $W_{-5,-3} = 1 - (|2|/(11 - 1)) = .8$
- $W_{-5,-2} = 1 - (|3|/(11 - 1)) = .7$
- $W_{-5,-1} = 1 - (|4|/(11 - 1)) = .6$

Table 22: Linear Weights

					Judge2						
Judge1	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**
-5	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
-4	0.9	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
-3	0.8	0.9	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
-2	0.7	0.8	0.9	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3
-1	0.6	0.7	0.8	0.9	1.0	0.9	0.8	0.7	0.6	0.5	0.4
0	0.5	0.6	0.7	0.8	0.9	1.0	0.9	0.8	0.7	0.6	0.5
1	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.9	0.8	0.7	0.6

Table 22: Linear Weights

	Judge2										
Judge1	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**
2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.9	0.8	0.7
3	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.9	0.8
4	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.9
5	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Notice that each concordant value is 1 and all values next to it are 0.9. This creates a cascade effect for the weights.

Quadratic weights, formally known as Fleiss-Cohen weights, use quadratic distancing between the weights. A cell's weight is quadratically related to its distance from the concordant value.

The formula for the quadratic weights are:

- $W_{ij} = 1 - (|i - j|)^2/(R - 1)^2$

Again, let's calculate the weights of the first few cells using the the formula. We'll begin with the (-5, -5) cell and move right on the table.

- $W_{-5,-5} = 1 - (|0|^2/(11 - 1)^2) = 1$
- $W_{-5,-4} = 1 - (|1|^2/(11 - 1)^2) = .99$
- $W_{-5,-3} = 1 - (|2|^2/(11 - 1)^2) = .96$
- $W_{-5,-2} = 1 - (|3|^2/(11 - 1)^2) = .91$
- $W_{-5,-1} = 1 - (|4|^2/(11 - 1)^2) = .84$

Table 23: Linear Weights

	Judge2										
Judge1	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**
-5	1.00	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19	0.00
-4	0.99	1.00	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19
-3	0.96	0.99	1.00	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36
-2	0.91	0.96	0.99	1.00	0.99	0.96	0.91	0.84	0.75	0.64	0.51
-1	0.84	0.91	0.96	0.99	1.00	0.99	0.96	0.91	0.84	0.75	0.64
0	0.75	0.84	0.91	0.96	0.99	1.00	0.99	0.96	0.91	0.84	0.75
1	0.64	0.75	0.84	0.91	0.96	0.99	1.00	0.99	0.96	0.91	0.84
2	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1.00	0.99	0.96	0.91
3	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1.00	0.99	0.96
4	0.19	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1.00	0.99
5	0.00	0.19	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1.00

Again, notice how each concordant value is 1 and all values next to it are .99. This creates a steeper cascade than the linear weighting as the differences in judging increase.

Assess the two weighting methods for yourself. What are the advantages and disadvantages of each? Can you imagine any problems arising for either? Which would you choose for our MMA data and why?

Linear weighting values the distance between the fourth and fifth category the same as the distance between the first and second category. If this constant effect fits the data, then it's best to choose linear weighting.

Quadratic weighting determines that the distance between the first and second category is much less than the distance between the fourth and fifth category. As the categories get furthered removed from the concordant value, the difference becomes more egregious.

For the MMA data, we tend to think the quadratic weighting method works best. Generally, egregious misses are the errors that cast doubt on the judging system. The difference in a 3 point and 2 point win is basically none. Still, we need to be careful. Under the quadratic weighting method, a 1 point win for fighter 1 and a 1 point win for fighter 2 are essentially in agreement (w = .96).

Calculating and Interpreting Weighted Kappa

The calculation and interpretation of the weighted kappa κ are the same as Cohen's kappa. If you need a refresher, read through our explanation in the previous tab.

Our weighted kappa (κ) is calculated once again by $kappa(\kappa) = \frac{p_o - p_e}{1 - p_e}$.

with weights:

- $p_o = \sum_i \sum_j W_{ij} P_{ij}$
- $p_e = \sum_i \sum_j W_{i+} P_{+j}$.

Using quadratic weights, the observed proportion of agreement is 0.982. This is extremely high, because we have so many partial agreements. If you're curious, look again through our contingency table.

However, the expected proportion of agreement is also very high at 0.9. The weights may inflate our observed agreement levels by adding in partial agreement, but they also inflate our expected agreement.

After solving for D'Amato and Lee's weighted kappa, we find it at 0.82. This is higher than the unweighted kappa (for the outcome variable) of 0.65, likely because a lot of judges disagree marginally.

Our interpretation for the weighted kappa is identical to that of Cohen's kappa. Below is a reminder:

- $\kappa > 0.75$ Excellent reproducibility
- $0.4 \leq \kappa \leq 0.75$ Good reproducibility
- $0 \leq \kappa < 0.4$ Marginal reproducibility

Using quadratic weights, we can say there is excellent reproducibility in the scoring margins between D'Amato and Lee. This means the judges are generally consistent in their scores and it is possible to replace one with the other and expect similar results. Remember, as with the other measures, this kappa does not mean that the judges are accurate in their assessments.

We calculate the confidence interval the same way as before, and our confidence interval is from 0.581 to 1.059. Thus, with 95% confidence, the interval includes the true kappa value for these judges.

This should give us pause. The lower end of our confidence interval is 0.581. This means the true kappa could be this low. This would drop our verdict to "good reproducibility" and change our overall assessment of the relationship.

Like Cohen's kappa, our kappa test has a null and alternative hypothesis of:

- $H_o : \kappa = 0$
- $H_a : \kappa > 0$

The confidence interval doesn't include 0, so we have sufficient evidence to reject the null hypothesis that there is no association between the judges' scores.

Comparing Kappa and the Weighted Kappa

Let's compare our results with the linear weights. The observed proportion of agreement is 0.919 and the expected proportion of agreement is 0.747. The linear-weighted kappa is 0.68.

This drops our interpretation to only "good reproducibility". We can be reasonably confident in the judges' reproducibility, but it's also feasible that swapping D'Amato for Lee could lead to a different result. An estimated 32% of the data is composed of chance agreement or disagreement. Once again, this would not indicate that D'Amato or Lee are somehow less accurate than before, it only speaks to their consistency and reproducibility.

We can say with 95% confidence that 0.486 and 0.874 contains the true kappa value. Once again, the interval does not include 0, so we have sufficient evidence to reject our null hypothesis that there is no association between the judges' rulings. The lower end of the interval is 0.486. This would indicate "good reproducibility". As with the quadratic weighting, this should lower our assessment of the relationship between the two judges.

Exercise 7: D'Amato and Cleary

Now that we've walked through an example of weighted kappa on the consistency of D'Amato's and Lee's scoring margins, let's look at the scoring margins of D'Amato and Cleary. We'll present the data and the findings to you, and you can reproduce the analysis. Feel free to look at our earlier phrasings and points.

Once again, we filtered the data to only include three rounds. We displayed all of the scoring margins by the judges in a contingency table below.

Table 24: D'Amato and Cleary's Decision Margin

						Cleary						
D'Amato	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
-5	2	1	0	0	0	0	0	0	0	0	0	**3**
-4	0	3	1	0	0	0	0	0	0	0	0	**4**
-3	0	1	17	0	3	0	0	0	0	0	0	**21**
-2	0	0	0	0	1	0	0	0	0	0	0	**1**
-1	0	0	3	1	20	0	5	0	0	0	0	**29**
0	0	0	0	0	0	3	0	0	0	0	0	**3**
1	0	0	0	0	2	0	20	0	2	1	0	**25**
2	0	0	0	0	0	0	0	1	0	1	0	**2**
3	0	0	0	0	0	0	7	0	14	0	0	**21**
4	0	0	0	0	0	0	0	2	1	2	1	**6**
5	0	0	0	0	0	0	0	0	0	1	1	**2**

Table 24: D'Amato and Cleary's Decision Margin

						Cleary					
D'Amato	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
Total	**2**	**5**	**21**	**1**	**26**	**3**	**32**	**3**	**17**	**5**	**2**	**117**

7.1. How often do D'Amato and Cleary completely agree?

SOLUTION: 83 times (71 percent)

7.2. Do D'Amato and Cleary seem to agree on the fight often but not on the score?

SOLUTION: Yes. There are only 7 times one is positive and the other negative.

7.3. How does D'Amato and Cleary agreement compare to D'Amato and Lee (previous exercise)?

SOLUTION: They agree more. (71 vs 58 percent)

Table 25: Expected Values of D'Amato and Cleary

						Cleary					
D'Amato	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
-5	0.1	0.1	0.5	0.0	0.7	0.1	0.8	0.1	0.4	0.1	0.1	**3**
-4	0.1	0.2	0.7	0.0	0.9	0.1	1.1	0.1	0.6	0.2	0.1	**4**
-3	0.4	0.9	3.8	0.2	4.7	0.5	5.7	0.5	3.1	0.9	0.4	**21**
-2	0.0	0.0	0.2	0.0	0.2	0.0	0.3	0.0	0.1	0.0	0.0	**1**
-1	0.5	1.2	5.2	0.2	6.4	0.7	7.9	0.7	4.2	1.2	0.5	**29**
0	0.1	0.1	0.5	0.0	0.7	0.1	0.8	0.1	0.4	0.1	0.1	**3**
1	0.4	1.1	4.5	0.2	5.6	0.6	6.8	0.6	3.6	1.1	0.4	**25**
2	0.0	0.1	0.4	0.0	0.4	0.1	0.5	0.1	0.3	0.1	0.0	**2**
3	0.4	0.9	3.8	0.2	4.7	0.5	5.7	0.5	3.1	0.9	0.4	**21**
4	0.1	0.3	1.1	0.1	1.3	0.2	1.6	0.2	0.9	0.3	0.1	**6**
5	0.0	0.1	0.4	0.0	0.4	0.1	0.5	0.1	0.3	0.1	0.0	**2**
Total	**2.0**	**5.0**	**21.0**	**1.0**	**26.0**	**3.0**	**32.0**	**3.0**	**17.0**	**5.0**	**2.0**	**117**

7.4. Look through the expected value table. Do the highly expected values also occur frequently in the observed table?

SOLUTION: Somewhat. The highest observed values have relatively high exepected, but some high expected values are not observed often.

Quadratic Weights:

- Proportion of Observed Agreement: 0.99
- Proportion of Expected Agreement: 0.878
- Weighted kappa with Quadratic Weights: 0.92

- 95% Confidence Interval for kappa: 0.772 and 1.068

7.5. Using quadratic weights do D'Amato and Cleary have good reproducibility?

SOLUTION: Yes, the kappa value is over 0.9.

7.6. (Optional - Advanced) Assess the confidence interval based on quadratic weights. What can you conclude?

SOLUTION: We have evidence to support excellent reproducibility for these two judges.

Linear Weights:

- Proportion of Observed Agreement: 0.948
- Proportion of Expected Agreement: 0.722
- Weighted kappa with Linear Weights: 0.81
- 95% Confidence Interval for kappa: 0.665 and 0.955

7.7. Using linear weights do D'Amato and Cleary have good reproducibility?

SOLUTION: Yes, the kappa value is over 0.9.

7.8. (Optional - Advanced) Assess the confidence interval based on linear weights. What can you conclude?

SOLUTION: We have too much variability so there is not enough evidence to conclude reproducability is excellent.

7.9. (Optional Advanced) How large is the difference between the quadratic and linear weights?

SOLUTION: A large enough difference to matter when drawing inference from the confidence interval.

Fleiss' Kappa

Thus far, we've assessed the inter-rater reliability within data sets of two judges, but what about three or more judges? MMA fights are evaluated by three judges, and in both the weighted and unweighted variations of Cohen's kappa, we completely ignore the third judge. This ignorance becomes even more egregious if we have larger quantities of judges.

Several different methodologies have been created to account for this. **Light's kappa**, for example, takes the average of every combination of Cohen's kappa within the pool of raters. We'll turn to a slightly more complex version. **Fleiss' kappa** is a variation of Cohen's kappa that allows for three or more judges. It measures the level of agreement or consistency within the group of judges. A high Fleiss' kappa would indicate a high rate of reliability between the group of judges.

Fleiss' kappa works with nominal variables. It does not give weight to partial agreement like weighted kappa. There are methods that work with ordinal variables and partial agreement with three or more judges, but they extend beyond the scope of this module. Search for **Kendall's Coefficient of Concordance** if you are interested.

Like all other kappa values, Fleiss' kappa removes chance agreement. Because the method is unweighted and gives out no partial agreement, we'll use the outcome variable for our analysis.

Fleiss's kappa makes a few assumptions about the data. They are similar to the assumptions made by weighted kappa and Cohen's kappa, but not exactly the same.

1. Each of the raters are **independent**.
2. The raters are selecting from the same defined categories of a categorical variable.

We've selected a new set of three judges from our data set that judged lots of fights together. Cartlidge, Collett, and Lethaby judged 96 fights that went to a decision together.

With three or more judges, it becomes difficult to observe the data using a contingency table.

Below is a table of each judge's verdict for the 96 fights. We've created three columns on the right to help summarize the judge's votes. They sum up the total number of verdicts of that type for each fight.

fight	Cartlidge	Collett	Lethaby	fighter2	draw	fighter1
1	fighter2	fighter2	fighter2	3	0	0
2	fighter1	fighter1	fighter1	0	0	3
3	fighter1	fighter1	fighter1	0	0	3
4	fighter2	fighter2	fighter2	3	0	0
5	fighter2	fighter2	fighter2	3	0	0
6	fighter1	fighter1	fighter1	0	0	3
7	fighter1	fighter1	fighter1	0	0	3
8	fighter1	fighter1	fighter1	0	0	3
9	fighter1	fighter1	fighter1	0	0	3
10	fighter1	fighter2	fighter2	2	0	1
11	fighter1	fighter1	fighter1	0	0	3
12	fighter2	fighter2	fighter2	3	0	0
13	fighter1	fighter1	fighter1	0	0	3

fight	Cartlidge	Collett	Lethaby	fighter2	draw	fighter1
14	fighter2	fighter2	fighter2	3	0	0
15	fighter1	fighter1	fighter1	0	0	3

Exercise 8: Cartlidge, Collett, and Lethaby

8.1. Take a look at the table. How often do the judges agree?

SOLUTION: All but one fight

Calculating and Interpreting Fleiss' Kappa

As with the other kappas, we begin by calculating the the proportion of observed agreement (p_o) and proportion of expected agreement (p_e). However, for Fleiss' kappa, they are calculated in more complex ways.

This makes sense. As we add more judges, we have so many more levels of agreement. For example, if Collett and Lethaby agree, but Cartlidge disagrees (like fight 10 in our data above), this is still better agreement than if all three judges give different verdicts. These options are only magnified if we were to consider sets of four or more judges or events with four of more different outcomes.

The proportion of observed agreement is calculated by a long formula:

- $p_o = \frac{1}{N * n * (n-1)}(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - N * n)$

where N is the number of observations and n is the number of raters.

For our example, N = 96 and n = 3.

You won't have to calculate it by hand, and in this case intuition for the formula is not easy and beyond the scope of the module.

We show the calculation using this formula for the proportion of observed agreement for our set of Cartlidge, Collett, and Lethaby:

- $p_o = \frac{1}{96 * 3 * (3-1)}(3^2 + 0^2 + 0^2 + ... + 1^2 - 96 * 3)$

(Optional) Take a moment if you are interested to see how we entered the values into the formula.

After evaluating, we end up with a $p_o = 0.882$. This is our total observed agreement. It includes both real agreement and chance agreement.

The proportion of expected agreement is computed by a much less complex formula.

- $p_e = \sum p_j^2$

We calculate the frequency (or expected rate) for each of the three categories (p_j), square them, and add them all together. This is like finding the concordant values with two judges. We're finding the probability that the selections appear together randomly.

- $p_{fighter1} = 0.524$

- $p_{draw} = 0.017$
- $p_{fighter2} = 0.458$

If we square these frequencies and sum them up, we'll find that $p_e = 0.485$.

This means that if the three judges were to issue random verdicts without watching the fights or consulting with each other, we'd expect the three of them to agree about 48.5% of the time.

We can solve for Fleiss' kappa (κ) with the same formula as the weighted and unweighted kappa values.

- $kappa(\kappa) = \frac{p_o - p_e}{1 - p_e}$.

Fleiss' kappa for Cartlidge, Collett, and Lethaby is 0.771.

Our interpretation for the Fleiss' kappa is identical to that of the weighted and unweighted kappa. Below is a reminder:

- $\kappa > 0.75$ Excellent reproducibility
- $0.4 \leq \kappa \leq 0.75$ Good reproducibility
- $0 \leq \kappa < 0.4$ Marginal reproducibility

We can claim that Cartlidge, Collett, and Lethaby have excellent reproducibility in their judgments. This means they are likely to evaluate the fights in similar ways, and if we substituted one for another, we would not expect exceedingly different results. About 23% of the data is a result of chance agreement or disagreement. Once again, this cannot prove that the three of them are good at selecting the correct victor, only that they are likely to select similar victors.

As with the other kappa values, we can calculate a confidence interval. Our 95% confidence interval for Fleiss' kappa is 0.661 to 0.881. Thus, with 95% confidence, we can claim that the interval includes the true value of Fleiss' kappa. This interval does not include 0, so we can conclude with at least 95% confidence that there is some real association between the three judges. The lower end of the confidence interval is 0.661, which would be in the upper portion of the "good reproducibility" bracket.

Fleiss' kappa does afford us an extra piece of analysis. We can look at the individual kappas for each of the categories to assess the level of agreement across their verdicts. This can help us to break down our kappa into simpler results that assess raters reliability on only one category.

This can be especially helpful for certain tests of reliability. For example, a survey evaluating the inter-rater-reliability of several doctors prescribing or diagnosing patients would immensely benefit by seeing which prescriptions or diagnoses the doctors are most and least consistent in the ratings.

For our data, we'll look at the individual kappas for the categories: fighter1, draw, and fighter2.

Category	Kappa	z	p.value
draw	0.186	3.154	0.002
fighter1	0.791	13.427	0.000
fighter2	0.790	13.410	0.000

Fighter 1 and fighter 2 are arbitrary assignments, so it is fitting that their values are almost identical. Their difference would not tell us anything meaningful regardless. However, the individual kappa of the draw category is much smaller than the others. This demonstrates that the judges have a much lower level of agreement when issuing draws than when they select a fighter.

This makes contextual sense. Draws are unlikely and less desirable. Collett, Cartlidge, and Lethaby never put forth a unanimous draw, and they rarely even had two of three vote draw.

Exercise 9: Other Judges

We'll provide a second example using the judge combination of Cartlidge, Sledge, and Lethaby. We'll ask you some general questions to help guide your analysis.

judge1	judge2	judge3	fights
Cartlidge	Collett	Lethaby	96
Cartlidge	Lethaby	Sledge	36
Champion	Divilbiss	Graham	33
Cleary	D'Amato	Lee	24
Crosby	D'Amato	Valel	22
Cleary	D'Amato	Kamijo	18
Cartlidge	Lethaby	Oglesby	14
Colón	Tirelli	Urso	14
Gueary	Miller	Swanberg	14
Mathisen	Sutherland	Turnage	14

fight	Cartlidge	Lethaby	Sledge	fighter2	draw	fighter1
1	fighter2	fighter2	fighter2	3	0	0
2	fighter2	fighter2	fighter2	3	0	0
3	fighter1	fighter2	fighter2	2	0	1
4	fighter2	fighter2	fighter2	3	0	0
5	fighter1	fighter1	fighter1	0	0	3
6	fighter2	fighter2	fighter2	3	0	0
7	fighter2	fighter2	fighter2	3	0	0
8	fighter1	fighter1	fighter1	0	0	3
9	fighter1	fighter2	fighter2	2	0	1
10	fighter1	fighter1	fighter1	0	0	3
11	fighter1	fighter1	fighter1	0	0	3
12	fighter1	fighter1	fighter1	0	0	3
13	fighter2	fighter2	fighter2	3	0	0
14	fighter2	fighter2	fighter2	3	0	0
15	fighter2	fighter1	fighter1	1	0	2

9.1. Take a look at the table. How often do the judges agree?

SOLUTION: Most fights (12 of 15)

9.2. Does one judge tend to differ more?

SOLUTION: Yes, Cartlidge is the only one to disagree

Results

- Proportion of Observed Agreement: 0.926
- Proportion of Expected Agreement: 0.529
- Fleiss' kappa: 0.843
- 95% Confidence Interval for kappa: 0.843 and 0.654

Table of Individual kappas:

Category	Kappa	z	p.value
fighter1	0.843	8.758	0
fighter2	0.843	8.758	0

9.3. Based on the results, what is the percent disagreement that is attributable to chance?

SOLUTION: Just over half (53 percent)

9.4. How good is the estimated reproducibility for the three judges?

SOLUTION: Excellent, well above chance agreement (over 90 percent)

9.5. (Optional advanced) Assess the confidence interval. Can you reject a null hypothesis of excellent reproducibility?

SOLUTION: No, the interval includes the possiblity of a value that is only good

Conclusion

In this module you have learned about metrics to measure consistency and reliability between raters and applied them to judges of MMA fights. In the examples and exercises, the overall consistency in judging was high but clearly higher for some judge pairs than others.

Recall our note about the data, which uses total scores instead of scores on each round. Our analysis likely produced higher reliability scores than if we used the individual round scores.

Analysis such as this for all judges could help inform the MMA about where, perhaps, they might spend time working with judges to improve the quality and consistency of scoring fights. The MMA may wish to consider these measures using data for each round. Such metrics would help them with judge selection and perhaps training to improve fairness.

Judges decisions can appear arbitrary and, in some cases, are controversial when there is disagreement. Fans and athletes alike benefit when the subjective nature of scoring is lessened.