To challenge or not to challenge?

In any sport, challenging a referee's decision is risky. The challenge process in American football is particularly chaotic, and coaches have just precious seconds in which to decide whether to throw the red challenge flag and gamble their team's fortunes. Can statistics help them? **Michael Orkin** investigates

During the course of an American National Football League (NFL) football game, the teams move up and down the field in a series of "downs" – short bursts of play – running or passing the ball in an attempt to score points by making either a "touchdown" for 6 points (plus a possible extra point by a short kick through the opponent's goal posts) or a field goal, a kick from the current position through the opponent's goal posts, for 3 points.

A running or passing play ends when the player with the ball is tackled by a player on the opposing team, at which point the teams generally have 40 seconds to regroup and run the next play (25 seconds if there is an administrative stoppage). The team with the ball has four downs, or attempts, to move the ball down the field. As soon as a team moves at least 10 yards farther down the field, they get a new set of four downs. If no points have been scored by the fourth down, the team with the ball must either "punt" (kick to the other team), attempt a field goal, or relinquish the ball.

During the gap between plays, the coach on either team can throw a weighted red flag onto the field to challenge the result of the play (this is only possible for certain types of plays, such as whether a pass was successfully caught or whether a player fumbled the ball). The coach must make this decision before the team with the ball runs another play, after which a challenge is no longer allowed. Each team is allowed three red flag challenges in each half of the game. The coach's decision on whether or not to throw the red flag is largely based on instant video replay, which is available after every play to both teams and even, to some extent, to the fans by way of the large video display in stadiums where the games are played. If the red flag challenge is rejected by the referees, the team that made the challenge is penalised by losing a timeout and play resumes. If the challenge is

Michael Orkin, PhD, is a mathematics professor at Berkeley City College and professor of statistics, emeritus, at California State University, East Bay.

accepted, there is no loss of timeout and the challenged play is negated. Since each team only has three timeouts per half, which last 2 minutes and are usually used for regrouping, they are important and the coaches do not want to waste them on a challenge they might lose. I will compare the red flag challenge to a bet in the casino game of craps and, in doing so, find a simple strategy for when to throw the red flag.

Example: 49ers v. Eagles

On 29 January 2023, in the first quarter of the National Football Conference (NFC) championship game between the San Francisco 49ers and the Philadelphia Eagles, the Eagles were on the fourth down with 3 yards to go – having had made 7 yards on the previous three downs – at the 49ers 35-yard line. The Eagles essentially had two choices: to -attempt a field goal for 3 points, or to try to move the ball the necessary 3 yards down the field with a run or a pass to get a "first down" (so named because if you successfully convert fourth down you get a first down), in which case the Eagles would get four more chances to score. They could also have "punted", but were too close to scoring to make punting sensible.

Rather than attempt a field goal, Philadelphia quarterback Jalen Hurts threw a 29-yard pass to Devonta Smith, who seemed to make a miraculous, one-handed catch at the 49ers 6-yard line, giving the Eagles "first and goal" (i.e., a first down only 6 yards from a touchdown). As soon as the play was over, Smith could be seen urging his teammates to run to the line of scrimmage (point of play) and begin the next play before 49ers coach Kyle Shanahan could throw the red flag and challenge the catch. This was a clear indication that Smith thought he had fumbled his catch and that the play would not stand up to the scrutiny of the instant replay. Additionally, the TV instant replay showed that Smith might have landed on the ground without full possession of the ball, in which case the pass would have been ruled incomplete. Despite these clues, Shanahan chose to not challenge the play, and the Eagles went on to score.

After the game, Shanahan said, "I wasn't going to throw [the challenge flag] just to hope and take a chance. That looked like a catch and we didn't want to waste a timeout … I heard they got a couple other angles later that (showed) it was not a catch" (bit.ly/3TDxDdF).

While timeouts are valuable, the possible loss of a timeout is often not as important as the value of a successful challenge. Also, since the 49ers had not taken any timeouts yet, even if they lost a challenge, they would still have had two timeouts remaining in the half. Since this was a fourth down play, if a challenge were successful, the pass would have been called incomplete, there would have been a change of possession and the 49ers would have had a first down at their own 35-yard line. If they lost the challenge, the Eagles would have had first and goal at the 49ers 6-yard line, which is exactly what happened when the 49ers did not challenge the play. A successful challenge would have completely changed the tempo of the game.

It turns out that if there is a reasonable chance of success, then the greater the "payoff" or value of a successful challenge, the more willing a coach should be to throw the red flag. Plays that can cause great damage, like the previous example, should be challenged more often than plays that would cause less harm. This can be quantified.

Instant replay history

According to data from an NFL rules website (bit.ly/3tdm8P0), in the 2021 and 2022 seasons an average of 56% of instant replay results, including challenges, resulted in the play being reversed. (In limited circumstances, like in the last 2 minutes of a game, the referees will analyse the instant replay results of a play and decide whether to reverse the play on their own, without a coach's challenge.)

If you go back to 1999, when modern instant replay began, 40% of instant replay results have resulted in reversals.

The casino game of craps

Let us compare the red flag challenge to the casino game of craps (dice). There are obvious differences, but in both situations if you make a bet (or a challenge), you either win or lose. Also, in both situations, there is a certain probability of winning, and there are payoff odds for winning. However, in a red flag challenge, the chance of winning and the payoff odds are estimated by the coach and his staff and vary from challenge to challenge, whereas in craps, these are fixed numbers.

In craps, you can bet $1 that a 7 will come up on the next roll of the dice. If you win, you are paid $4 in profit. If you lose, you lose the dollar that you wagered. In other words, the payoff odds set by the casino are 4 to 1. The chance that you win (chance of rolling a 7) equals 1/6 and the chance you lose equals 5/6. By the law of averages, if you make this bet repeatedly, betting $1 each time, you'll win $4 about 1/6th of the time and lose $1 about 5/6th of the time, for a long-run average loss of $4 \times \frac{1}{6} - 1 \times \frac{5}{6} = -\frac{1}{6} = -\$0.17$, about 17 cents per dollar bet.

This number is called the "expectation". Bets like this with negative expectation guarantee a long-run loss for the persistent gambler and a long-run profit for the casino. On the other hand, a bet with positive expectation can guarantee that a prudent gambler will make a long-run profit. For example, liberal blackjack rules provide a game with positive expectation for players who learn to count cards and who are playing under ideal conditions.

You can change a bet like a craps bet to a bet with positive expectation by modifying the probability that you win, which we will denote by p. Unsavoury craps players use "loaded" dice to try and achieve this. First, you find the break-even point, for which expectation equals 0, by solving the following expectation equation for p, with payoff odds denoted by K:

Expectation $= K \times p - (1 - p) = 0$

which becomes

$$p = \frac{1}{K+1}$$

For the modified bet on 7 in craps, with payoff odds 4 to 1, the win probability for the break-even point is

$$p = \frac{1}{5}$$

If your chance of winning equals 1/5, then the payoff odds of 4 to 1 make the expectation equal 0. For any win probability higher than 1/5 the bet has positive expectation, which is what cheats who use loaded dice try to accomplish.

It follows that for any "win–lose" game like craps, with win probability equal to p, and payoff odds of K to 1, the game has positive expectation whenever $p > 1/(K + 1)$.

The red flag challenge strategy

The craps bet and its possible modification suggests a strategy for deciding whether to throw the red flag and challenge a play in an NFL game. The coach must estimate two numbers: the chance of a successful challenge, p, and the value of a successful challenge, K (payoff odds). Here's the strategy: throw the red flag whenever $p > 1/(K + 1)$. This makes the red flag challenge similar to a win-lose game with positive expectation.

Applying the formula

Estimating the two numbers, p and K, can be straightforward, especially for experts like a coach and his staff. If you start with K, the payoff, you can see the range of p needed for a red flag challenge, and you can make a decision from there (or you can start with p and estimate K). Estimating p relies on available instant replay, with historical data as a guideline. If instant replay shows a receiver failing to catch the ball as he runs out of bounds, p might be at least 0.4, perhaps higher, given that in the past 2 years, instant replays resulted in reversals about 56% of the time. If instant replays and visual observations show no evidence of a problem, p may be very low, and the red flag should not be thrown, depending on the "payoff odds", that is, the value of a successful challenge.

Suppose you think the payoff odds for successfully challenging a play are 2 to 1. This means that the value to you of a reversal is twice the value of the referees' letting the play stand and your losing a timeout. In this case, throw the red flag if you think you have a better than $1/(1 + 2) = 1/3$ chance of a reversal (not unreasonable, considering the historical rate of instant replay reversals and what you may have seen in an instant replay). Here's a table with break-even points for different payoff values. For a given payoff, K, the strategy is to throw the red flag if your estimated chance of winning the challenge is greater than Prob(win) in the table.

Value = K	Prob(win)
5	1/6
4	1/5
3	1/4
2	1/3
1	1/2
0.5	2/3

Remember that throwing the red flag is not a casino game (although I could see it made into an app game). The payoff odds and probability of success are estimated numbers. Also, things change from challenge to challenge, so the law of averages cannot be applied. Still, a mathematical model does not have to be exactly the same as the real-life situation being modelled. By comparing the red flag challenge to a craps game, you can gain insights, including a strategy for when to throw the red flag, as in the table above.

Although K and p are estimated numbers, the relationship between K and p shows that as your value (payoff odds) for a successful challenge increases, so should your confidence in throwing the red flag, because the estimated probability needed for positive expectation decreases. If there's not much to gain by a challenge, then you should be more conservative about throwing the red flag. The fact that 56% of instant replay decisions have resulted in reversals in the past 2 years can be used to help determine your estimated probability of success when making this split-second judgement.

Back to 49ers v. Eagles

In the 49ers v. Eagles game, there was a huge payoff for winning a challenge to Devonta Smith's completed pass, including the Eagles' loss of possession due to a turnover on downs. Either the Eagles would have had the ball with first and 10 on the 49ers 6-yard line (terrible) or the 49ers would regain possession of the ball on their own 35 (excellent). Suppose the coach thought that the value of a successful challenge was 3. Then the threshold probability of success was 1/4, or 0.25. Or perhaps the coach thought that the value of the challenge was 4. Then the threshold probability was 1/5. In any event, the threshold probability would have been below the recent average instant replay reversal rate of 56%, and, given the available instant replay information, any estimated success probability p was likely to be larger than 1/3. The model for this bet yielded positive expectation. Conclusion: Coach Shanahan should have thrown the red flag.

The 2023 Super Bowl

In the 2023 Super Bowl, with 6 minutes remaining in the third quarter, the Philadelphia Eagles had the ball, third and 14, at the Kansas City 47-yard line. Jalen Hurts threw a pass down the right sideline to tight end Dallas Goedert, who caught the pass at the Kansas City 30-yard line, but appeared to fumble the ball as he was going out of bounds. Kansas City coach Andy Reid threw the red flag, but upon further review, the referees ruled that the play should stand. As a result, Kansas City lost a timeout and Philadelphia had the ball first and 10 at the Kansas City 30.

This play was similar to the Eagles' play against the 49ers in the NFC championship game; however, since it was third down, the Eagles would have retained possession of the ball even if the referees had reversed the play. Thus, Coach Reid's payoff odds would not have been as high as the play in the 49ers game but still would have been high, since the result of the play would have given Philadelphia a first down in scoring position. Also, the fact that Goedert was fumbling the ball just before he stepped out of bounds made the probability of a successful challenge reasonably high (whether or not the referees got it right is still being debated). It appears that Coach Reid correctly challenged the play, even though his challenge was unsuccessful. Subsequently, the Chiefs were resilient and kept the Eagles to a field goal on their way to winning the Super Bowl.

You may have noticed that both examples involved the Philadelphia Eagles, whose quarterback is Jaylen Hurts. It has been observed that Hurts may throw the ball harder and in riskier situations than many NFL quarterbacks, resulting in more uncaught balls than average. Coaches should be on the lookout for such challenge opportunities.

Conclusion

By comparing a coach's red flag challenge to the casino game of craps, there is a simple model that helps determine the parameters for making a challenge and that provides the coach and his staff with a straightforward strategy for when to throw the red flag. If the coach's estimates are good, the strategy provides a gamble with positive expectation. This not only provides the coaches with a simple guideline for making a red flag challenge, but also indirectly increases the fans' enjoyment of watching a game by knowing that the coaches have a tool to help improve the accuracy of the outcome. ∎