Glossary for *Statistical Reasoning in Sports*

An athlete's ***ABILITY*** is a true but unknown value that describes what the player would do if given an infinite number of opportunities in the same context.

The **alternative hypothesis**, denoted H_a, describes what we want to establish or what we suspect is true.

The **area principle** says that the area representing each category in a graph should be proportional to the number of observations in that category.

In basketball and hockey, a player is credited with an **assist** when his or her pass to a teammate leads to the teammate scoring a basket or goal.

Two variables have an **association** if specific values of one variable tend to occur in common with specific values of the other variable.

In golf, the **average driving distance** is how far a golfer hits the ball, on average, when using his or her driver.

A **bar chart** displays the possible outcomes of a categorical variable as individual equally wide bars, with the height of each bar proportional to how often each corresponding outcome occurred.

In an exponential model of the form $\hat{y} = a(b)^x$, b is called the **base**. The base determines the direction and steepness of the exponential model.

In baseball or softball, a player's **batting average** is the proportion of at-bats that the player hits the ball and safely makes it on base.

In baseball or softball, a player's **batting average on balls in play (BABIP)** measures the proportion of times that he or she hits the ball into the field of play and reaches base. This excludes home runs, strikeouts, and walks. A typical BABIP is about 0.300:

$$\text{BABIP} = \frac{\text{Hits - Home Runs}}{\text{Plate Appearances - Home Runs - Strikeouts - Walks}}$$

In baseball and softball, the nine hitters in the game bat in a specific **batting order** determined before the game begins. Once the batting order is turned in to the umpires, the order cannot change, although a particular player may be replaced by a different player in the same position in the batting order.

In an experiment, a subject is **blind** if he or she does not know which treatment he or she is receiving. This prevents the subject from consciously or subconsciously altering his or her response. It is also possible, and usually beneficial, if the people collecting the data are also blind. If subjects and the people collecting the data are blind, the experiment is called double blind.

A **boxplot** is a visual representation of the five-number summary of a numerical distribution.

An athlete is considered to be a **clutch** player if his or her *ABILITY* to be successful is higher during important situations in a game.

Variables are **categorical** if the possible outcomes fall into categories. Typically, this means that the possible outcomes are described with words, such as "win" or "loss."

When every member of the population is studied, it is called a **census**.

In an exponential model of the form $\hat{y} = a(b)^x$, a is called the **coefficient**. The coefficient is the predicted value of $\hat{y}$ when $x = 0$.

Combinations are unordered collections of objects in which each possible object occurs at most once, but not all possible objects need to be used. The total number of combinations when choosing r objects out of n total choices is ${}_n\mathrm{C}_r = \frac{n!}{r!(n-r)!}$.

A **confidence interval** provides an interval of plausible values for an athlete's *ABILITY* or an interval of plausible values for the difference in an athlete's *ABILITY* in two different contexts.

The **confidence level** of a confidence interval describes how much confidence we should have that the interval of plausible values actually contains an athlete's true *ABILITY*. For example, if intervals were calculated for many athletes at the 95% confidence level, about 95% of the intervals would successfully contain the *ABILITY* of the athlete for whom the interval was calculated.

A **conditional probability** describes the probability that an event occurs, given that we know that a different event has occurred. Conditional probabilities are expressed in the form $P(A \mid B)$, which is read "The probability that event A occurs, given that event B has occurred."

Variables are **confounded** when we aren't certain which variable is causing the effect we have observed. For example, if a team always wears red jerseys at home games, the color of the jerseys and the comfort of playing at home are confounded, since we wouldn't be able to tell which variable is causing the team to play better at home.

In an experiment, **control** means keeping the conditions exactly the same, except for the treatments being compared.

An athlete's or team's *PERFORMANCES* are **consistent** if they do not vary much from the athlete or team's typical *PERFORMANCE*.

The **correlation (*r*)** is a measure of the strength and direction of a linear association between two numerical variables.

In baseball, the **designated hitter (DH)** is a player who plays on offense but not on defense, usually replacing the pitcher in the batting order.

A **deviation** measures the distance between an observed *PERFORMANCE* and the mean of a distribution: Deviation = *PERFORMANCE* – mean.

A **distribution** of a categorical or numerical variable identifies the possible outcomes of a variable and how often it takes those outcomes.

A **dotplot** shows the distribution of a numerical variable. The horizontal axis shows the possible outcomes of the variable and the number of dots above each outcome shows how often that outcome occurred.

In baseball or softball, a **double** is when a baseball player makes it to second base on a hit.

In golf, a **driver** is the club that golfers use when they want to hit the ball as far as possible. This club is almost exclusively used for the first shot on a particular hole.

In golf, **driving accuracy** is the percentage of times a golfer's tee shot lands in the fairway.

In baseball, a pitcher's **Earned Run Average** (**ERA**) measures how many earned runs the pitcher allows per 9 innings. Unearned runs, which are runs that score because of fielding errors, do not count in a pitcher's ERA.

$$\text{ERA} = \frac{\text{Number of Earned Runs}}{\text{Number of Innings Pitched}} \times 9$$

ERA+ adjusts a pitcher's ERA according to the pitcher's ballpark (in case the ballpark favors batters or pitchers) and the ERA of the pitcher's league. Average ERA+ is set at 100; a score above 100 indicates that the pitcher performed better than average, below 100 indicates worse than average.

In a chance process, an **event** is a subset of the sample space.

The **expected value** (mean value) of a random variable *X* is equal to the average value of the random variable if the chance process was repeated many, many times. To calculate the expected value, use the following formula: $\mu_x = E(X) = \Sigma X \cdot P(X)$. In a binomial distribution with *n* trials and *p* probability of success, the expected number of successes is $\mu_x = E(X) = np$.

An **experiment** deliberately imposes treatments on individuals to measure their responses.

In an experiment, the **explanatory variable** is what is deliberately changed to see if this change causes a change in the response variable. When displaying the relationship between two numerical variables on a scatterplot, the explanatory variable is the variable that we use to predict the value of the response variable. The explanatory variable is plotted on the horizontal axis (*x* axis).

An **exponential model** is a model in the form $\hat{y} = a(b)^x$ that is used to model a nonlinear association on a scatterplot. If $b > 1$, then the model is called an exponential growth model. If $b < 1$, then the model is called an exponential decay model.

Extrapolation is when we predict the value of the response variable for a value of the explanatory variable that is much larger or much smaller than the other values of the explanatory variable in the data set. It is risky to make predictions using extrapolation since the association between the variables may not be the same for extremely small or extremely large values of the explanatory variable.

The **factorial** of a positive integer *n*, denoted ***n*!**, is the product of the integer *n* and all the positive integers below it. For example, $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$. By definition, 0! = 1.

In golf, the **fairway** is an open path of relatively short grass that goes from the tee box to the hole. It is surrounded by the rough, which is much longer grass and other vegetation.

The **five-number summary** consists of the minimum, Q_1, median, Q_3, and maximum values in a distribution of numerical data.

The **frequency** describes the number of observations in each class or category.

The **fundamental counting principal** says that if one event can occur in *m* ways and a second event can occur in *n* ways for any of the occurrences of the first event, then the first event and the second event can occur in $m \times n$ ways.

In many sports leagues, the championship is determined by having two teams play a 7-game series, where the first team to win 4 games is the champion. If the teams each win 3 of the first 6 games, then the championship is determined by **game 7**.

The **general addition rule** says that for any two events *A* and *B*, *P*(*A* or *B*) = *P*(*A*) + *P*(*B*) - *P*(*A* and *B*).

In golf, to hit the **green-in-regulation** (GIR) means that you get the ball on the putting surface (the green) in at least two shots less than par.

A **histogram** divides the values of a numerical variable into classes and uses bars to represent the number of values in each class.

A **hit** is when a baseball or softball player hits the ball and makes it safely on base.

In baseball or softball, a **home run** is when a player hits a fair ball over the outfield fence. When this happens, the hitter and any runners on base automatically score a run. Very rarely a player hits an inside-the-park home run. In this case, the hitter is able to run all the way around the bases before the defense tags him out.

In baseball, a pitcher's **home run rate** is the average number of home runs he gives up per 9 innings.

A player has the **hot hand** if his or her *ABILITY* to be successful is higher following a success than following a failure. When a player has the hot hand, the percentage of successes following a success will be higher than the percentage of successes following a failure.

A **hypothesis test** is the formal process used to decide between two competing hypotheses about an athlete or team's *ABILITY*, called the null hypothesis and the alternative hypothesis.

Two events are **independent** if knowing the outcome of one event does not affect the probability of the other event. If $P(A \mid B) = P(A \mid \textit{not } B) = P(A)$, then events A and B are independent. An athlete's attempts are independent if his or her *ABILITY* to be successful is the same following a success and following a failure.

An **indicator variable** is a categorical variable with two possible outcomes. These outcomes are coded numerically so they can be included in regression calculations. Typically, a success is reported as a "1" and a failure is recorded as a "0."

The **interquartile range (*IQR*)** is a single number that measures the range of the middle 50% of the distribution. $IQR = Q_3 - Q_1$.

In golf, an **iron** is a club that golfers use when they want to hit the ball a specific distance. Typically, golfers will use 7 to 10 different iron clubs, each designed to hit the ball a different distance. For example, a 3-iron will hit the ball farther than a 7-iron, which will hit the ball farther than a 9-iron.

In a sports context, the **law of large numbers** says that an athlete's *PERFORMANCE* will generally get closer and closer to his or her *ABILITY* as the number of attempts grows larger.

A **least-squares regression line** is used to model a linear relationship between an explanatory variable x and a response variable y. It is usually expressed in the form $\hat{y} = a + bx$ where $\hat{y}$ ("y-hat") = the predicted value of y, a = the y intercept (constant), and b = the slope.

The form of an association on a scatterplot is **linear** if the pattern of the points is best described by a straight line.

A **logistic model** is a model in the form $\hat{p} = \frac{e^{a+bx}}{1+e^{a+bx}}$. Logistic models are used to predict the outcome of a *categorical* variable for a particular value of a numerical explanatory variable. The value of $\hat{p}$ in a logistic model will always be between 0 and 1, and it can be interpreted in two ways. If many observations are made at a particular value of *x*, then $\hat{p}$ is the prediction proportion of successes. If a single observation is made at a particular value of *x*, then $\hat{p}$ is the estimated probability of success.

The **margin of error** is added to and subtracted from a single-value estimate to create a confidence interval with the desired level of confidence. If 95% confidence intervals were calculated for many athletes, the distance between each athlete's actual *ABILITY* and his or her observed *PERFORMANCE* will be less than the margin of error in about 95% of the intervals.

The **mean** ($\bar{x}$) of a data set is its average value. To find the mean, add up all the values and then divide by the number of values in the data set. On a histogram or dotplot, the mean can be estimated by locating the balancing point of the distribution.

The **mean absolute deviation (MAD)** measures the average distance the values in a distribution are from their mean.

When we are using a numerical variable to compare the *ABILITIES* of athletes or teams in two different contexts and the data are paired, the test statistic we will use is the **mean difference**. To calculate the mean difference, find the difference in each pair and then calculate the mean of these differences. If the *ABILITIES* of the athletes or teams are the same in both contexts, the mean difference should be 0.

The **median** (*M*) of a data set is the middle value when the values are in order from smallest to largest. If there are two middle values, then the median is the average of the two middle values.

A **moving average** is the average of an athlete's *PERFORMANCES* in a specified time period and the time periods immediately before and after the specified time period.

A **multiple regression model** in the form $\hat{y} = a + b_1x_1 + b_2x_2 + \cdots$ uses more than one explanatory variable to predict the value of an explanatory variable. The explanatory variables can be numerical or categorical, but the response variable must be numerical.

Events *A* and *B* are **mutually exclusive** if they cannot occur together. That is, *A* and *B* are mutually exclusive events if they share no outcomes.

Two variables have a **negative association** if large values of one variable are typically paired with small values of the other variable.

Two variables have **no association** if knowing the values of one variable does not give any useful information about the values of the other variable.

The form of an association on a scatterplot is **nonlinear** if the pattern of the points is not best described by a straight line.

The graph of a Normal distribution is called a **Normal curve**. All Normal curves are symmetric, unimodal, and bell-shaped. The total area under a Normal curve is equal to 1 and the expected proportion of *PERFORMANCES* between two values is equal to the area under the Normal curve between these two values.

The **Normal distribution** is a mathematical model that is often used to describe distributions of data that are symmetric, unimodal, and bell-shaped.

The **null hypothesis**, denoted H_0, describes an initial belief that there has been no change in *ABILITY* or that there is no difference in *ABILITY* in two different contexts.

Variables are **numerical** if possible outcomes take on numerical values that represent different quantities of the variable.

A study that uses available data and does not impose treatments is called an **observational study**. Even when the data from an observational study provide convincing evidence that there really is a difference in *ABILITY*, it is unwise to conclude that changes in one variable *cause* changes in the other variable. This is because other variables are not controlled in an observational study and these variables might be confounded with the explanatory variable.

The **observed correlation** between two numerical variables is the correlation calculated from a limited amount of data, such as one season. The observed correlation will vary from the true correlation because of *RANDOM CHANCE*.

The **observed slope** of a least-squares regression line is the slope calculated from a limited amount of data, such as one season. The observed slope will vary from the true slope because of *RANDOM CHANCE*.

The **observed standard deviation** is the standard deviation calculated from a limited amount of data, such as one season. The observed standard deviation will vary from the true standard deviation because of *RANDOM CHANCE*.

In baseball and softball, a hitter's **on-base percentage** is the percentage of plate appearances where the batter gets on base.

In football, an **onside kick** occurs when a team kicking off deliberately kicks the ball a short distance, hoping to recover the football before the receiving team can gain possession of the football. To be eligible to recover the kickoff, the kicking team must kick the ball at least 10 yards.

In baseball and softball, the **opponent's batting average** measures the batting average of hitters when facing a particular pitcher. Pitchers that are more successful will have lower values for this variable.

In baseball and softball, **OPS** is a hitter's on-base percentage plus his slugging average. It was developed to measure a hitter's overall offensive *PERFORMANCE*.

An **outlier** is any value that falls out of the pattern of the rest of the data. Outliers can have a big effect on some summary statistics, such as the mean, range, and standard deviation.

In reporting a team's **overall record,** it is traditional to put the number of wins first followed by the number of losses (followed by the number of ties, if there are any). For example, if a baseball team's record is 100-62, they had 100 wins and 62 losses.

A ***p*-value** describes how likely it is to get a test statistic at least as extreme as the observed test statistic by *RANDOM CHANCE,* assuming that the null hypothesis is true.

In golf, **par** is the expected number of shots it should take to get the ball into the hole from the tee box.

In experiments where each subject gets both treatments, the data are **paired** because there are two values of the response variable for each subject. In studies that use available data about a group of athletes or teams, data are **paired** when there are two observations from each athlete or each team.

A quarterback's **passer rating** is a measure of his *PERFORMANCE* that is based on his completion percentage, passing yardage, touchdowns, and interceptions. The highest possible rating is 158.3.

In baseball or softball, a pitcher throws a **perfect game** when he or she doesn't allow any hitters to reach base for the entire game. In other words, every single batter who faces the pitcher makes an out.

An athlete's ***PERFORMANCE*** is an observed value that describes what the athlete actually did in a specific context. We can use an athlete's *PERFORMANCE* to estimate his or her *ABILITY*.

Permutations are ordered sequences of objects in which each possible object occurs at most once, but not all possible objects need to be used. The total number of permutations when selecting r objects out of n total choices is ${}_n\mathrm{P}_r = \frac{n!}{(n-r)!}$.

In basketball, a player is called for a **personal foul** when the player makes illegal physical contact with a player on the other team.

A **pie chart** displays the possible outcomes of a categorical variable as slices of a circular pie, with the area of each slice proportional to how often each corresponding outcome occurred.

In baseball and softball, each time a player takes his or her turn at bat, it is called a **plate appearance**.

In hockey, individual players are awarded a **point** for each goal that they score and for each assist that they make. An assist is a pass that directly leads to a goal for one of the player's teammates.

A **population** is the collection of individuals that you want to know about.

Two variables have a **positive association** if large values of one variable are typically paired with large values of the other variable and small values are paired with small values.

In hockey, when a player is sent to the penalty box to serve a penalty, he or she cannot be replaced by another player on the team. The advantage gained by the opposing team is called a **power play**.

In a chance process, the **probability** of an event is the event's long-run relative frequency. In other words, the probability of an event describes how often the event will occur in repeated trials of a chance process.

A **probability distribution** lists the possible values of a random variable and how likely they are to occur.

In golf, **putting average** is the average number of putts a player takes to complete an 18-hole course.

A **quadratic model** is a model in the form $\hat{y} = ax^2 + bx + c$ that is used to model a nonlinear association on a scatterplot. The graph of a quadratic model is called a parabola. If $a > 0$, then the parabola opens up. If $a < 0$, then the parabola opens down.

The **quartiles** of a distribution divide the distribution into four groups of roughly equal size. The first quartile Q_1 separates the lowest 25% of the values from the upper 75%. The second quartile, also called the median, separates the lower 50% of the values from the upper 50%. The third quartile Q_3 separates the lowest 75% of the values from the upper 25%.

RANDOM CHANCE describes the variation between an athlete's *PERFORMANCES* and his or her *ABILITY*.

A **random variable** takes on numerical values that describe the outcomes of a chance process.

In an experiment, **randomly assignment** is used to assign the treatments to individuals so that no treatment is given an advantage. In an experiment involving only one individual, the order that the treatments are used must be determined using a chance process.

The **range** is a single number that measures the distance between the minimum value and the maximum value of a distribution.

When measuring the same variable in two different time periods, the tendency for better *PERFORMANCES* to follow poor *PERFORMANCES* and for worse *PERFORMANCES* to follow good *PERFORMANCES* is called **regression to the mean.**

A **relative frequency** histogram or bar chart shows the proportion (or percentage) of observations in each class or category rather than the number of observations.

In an experiment, **replication** means making sure that each treatment has an adequate number of trials so that any difference in the effect of the treatments can be identified.

A **residual** is the difference between the actual value of a variable and the predicted value of that variable: residual = actual value – predicted value. When a residual is positive, the actual value is greater than the predicted value, so the prediction was too low. When a residual is negative, the actual value is below the predicted value, so the predicted value was too high.

A **residual plot** displays the values of the explanatory variable on the horizontal axis and the values of the residuals on the vertical axis. If there is only random scatter left over in the residual plot, then the model has the same form as the association. However, if there is a pattern left over in the residual plot, then the model does not have the same form as the association.

A measure of center or spread is **resistant** if it isn't influenced by unusually high or unusually low values in a distribution.

In an experiment, the **response variable** measures the outcome of interest. When displaying the relationship between two numerical variables on a scatterplot, we are usually interested in predicting the value of the response variable. The response variable is plotted on the horizontal axis (*y* axis).

In golf, the rough is the longer grass and other vegetation that surrounds the fairway.

In baseball or softball, a batter is credited with a **run batted in (RBI)** when a runner scores as a direct result of his or her at-bat.

In baseball, a team's **run expectancy** in a particular situation is the average number of additional runs that the team would score if they could keep playing in that context over and over.

A **sample** is a subset of the population that is chosen for study. Information from a sample is typically used to make generalizations about the population from which the sample was drawn.

The **sample size** is equal to the number of observations (e.g., games, shots, at-bats) that we are analyzing. When the sample size is larger, the effects of *RANDOM CHANCE* are balanced out and a team or athlete's *PERFORMANCE* will be closer to their *ABILITY*.

In a chance process, the **sample space** is the set of all possible outcomes.

In hockey or soccer, a goalie's **save percentage** is the percentage of shots taken by the opposing team that are prevented from going into the goal. For example, if opposing teams took 100 shots and a goalie stopped 90 of them, then his or her save percentage is 90/100 = 0.90 = 90%.

A **scatterplot** is a graph that displays the relationship between two numerical variables.

In basketball, a player's **scoring average** is the mean number of points he or she scores per game. In golf, a player's scoring average is the average number of shots it takes the player to complete 18 holes. Remember that lower scores are better in golf!

The **significance level** of a hypothesis test is a predetermined level of evidence that is required to confidently rule out *RANDOM CHANCE* as a plausible explanation. For example, if the significance level is 5%, then the *p*-value must be less than 5% to reject the null hypothesis and have convincing evidence to support the alternative hypothesis. The significance level is sometimes called the "alpha level" and is denoted with the Greek letter alpha (e.g. $\alpha = 5\%$).

A **single-value estimate** is a single number that represents our best guess for an athlete's *ABILITY*. It is equal to the athlete's *PERFORMANCE* and is in the exact middle of a confidence interval for the athlete's *ABILITY*.

Slugging average, also called slugging percentage, is the average number of bases that a player gets per at-bat. Singles count for one base, doubles for two bases, triples for three bases, and home runs for four bases.

The **standard deviation** measures the variability (spread) in a distribution of numerical data using the squared deviations from the mean. It measures the typical distance between an athlete's *PERFORMANCES* and his or her *ABILITY*.

The **standard deviation of the residuals *s*** is an estimate of the typical distance between the actual values of a response variable and their corresponding predicted values.

The evidence in a hypothesis test is **statistically significant** if *RANDOM CHANCE* is not a plausible explanation for observed *PERFORMANCE*(*S*). Thus, whenever we reject the null hypothesis in favor of the alternative hypothesis, the evidence is statistically significant.

In a scatterplot, the **strength** of an association describes the amount of scatter there is from the overall form of the data. In a strong association, there isn't much scatter and predictions of the response variable will be fairly precise.

In baseball and softball, a **strikeout** is when a batter makes an out by getting three strikes when he or she is batting.

In baseball, a pitcher's **strikeout rate** is the number of strikeouts he gets per 9 innings, on average.

The **sum of squared residuals (SSR)** is one way to measure how well a model makes predictions. If a model makes good predictions, the difference between the actual and predicted values will be small, resulting in a small sum of squared residuals: $SSR = \Sigma(\text{actual value - predicted value})^2$.

In golf, a **tee** is used to elevate the ball slightly off of the grass, making it easier to hit the ball cleanly.

A **test statistic** is a measure calculated from a team's (or player's) *PERFORMANCES* that is used as evidence in a hypothesis test.

In basketball, when a player attempts a **three-point shot**, he or she shoots the ball from behind the three-point line and is awarded 3 points instead of the usual 2 points if the shot is made.

A **time plot** of a numerical variable plots each *PERFORMANCE* against the time at which it was measured in order to observe possible trends over time and departures from these trends. The time periods are placed on the horizontal axis of the graph and the variable being investigated is placed on the vertical axis. Connecting the points with line segments helps emphasize any change over time.

A **treatment** is what each individual in the experiment is assigned to do.

In baseball or softball, a **triple** is when a player hits the ball and makes it safely to third base. Typically, a "triples hitter" needs lots of speed to make it all the way to third base on a hit.

The **true correlation** between two numerical variables, like the *ABILITY* of an athlete, exists only in theory. It is the correlation between the two variables after an infinite number of observations. The observed correlation is an estimate of the true correlation, but will vary due to *RANDOM CHANCE*.

The **true slope** of a least-squares regression line, like the *ABILITY* of an athlete, exists only in theory. It is the slope of the least-squares regression line used to model the relationship between two variables after an infinite number of observations. The observed slope is an estimate of the true slope, but will vary due to *RANDOM CHANCE*.

The **true standard deviation**, like the *ABILITY* of an athlete, exists only in theory. It is the standard deviation of a distribution after an infinite number of observations. The observed standard deviation is an estimate of the true standard deviation, but will vary due to *RANDOM CHANCE*.

A **two-way table** displays the distribution of a categorical variable in at least two different contexts or the relationship between two categorical variables.

A **Type I error** occurs when we reject the null hypothesis and support the alternative hypothesis when *in reality* the null hypothesis is true.

A **Type II error** occurs when we fail to reject the null hypothesis and do not decide to support the alternative hypothesis when *in reality* the alternative hypothesis is true.

A **variable** is a characteristic or attribute of an athletic *PERFORMANCE*. A variable can take on different values for different *PERFORMANCES*.

When using a quadratic model, the **vertex** of a parabola is where the direction of the graph changes from positive to negative or from negative to positive. In either case, the x coordinate of the vertex can be found with the following equation: $x = \dfrac{-b}{2a}$.

A **standardized score** or ***z*-score** measures how many standard deviations a *PERFORMANCE* is above or below the mean: $z = \dfrac{PERFORMANCE - \text{mean}}{\text{standard deviation}}$. When a z-score is positive, the *PERFORMANCE* is above the mean; when a z-score is negative, the *PERFORMANCE* is below the mean.

To find a team's **winning percentage,** divide the number of wins by the total number of games and multiply by 100. For a team with 100 wins and 62 losses, winning percentage = $\dfrac{100}{100+62} = 0.617 = 61.7\%$.

A **walk** is when a batter is awarded first base after the pitcher throws four pitches out of the strike zone.

A pitcher's **walk rate** is the average number of walks he gives up per 9 innings.

In baseball, the acronym **WHIP** stands for Walks plus Hits per Inning Pitched, or $\text{WHIP} = \frac{\text{walks + hits}}{\text{innings pitched}}$. Because pitchers are trying to prevent hitters from reaching base, low values of WHIP indicate better pitching *PERFORMANCES*.

A team's **win probability** measures the proportion of games a team would win if they could replay the game over and over again in the same context.