Choosing a Single Best Guess: Measures of Central Tendency

Learning Objectives:

* Learn the definitions of the mean, median, mode, and midrange as it relates to reasonable expectation
* Learn how the mean, median, mode, and midrange can be derived as solutions to variational problems

Outline:

* The point of this blog: guessing well and reasonable expectation
* The point of this post: explore measures of central tendency as reasonable expectations and as cost minimizers
* The guessing game setup
* The mean as a probability weighted sum
* The median as 50% probability point
* The mode as the least surprising
* How to choose?
* Cost functions
* inequality cost
* absolute difference cost
* squared difference cost
* max difference cost
* Beyond point estimation: distribution estimation
* The rules of the game define what is the best guess: w/wo replacement sampling
* The arbitrariness of cost functions
* Uncertainty in the distribution

Outline version 2.0

* Introduction
  + The purpose of the blog
  + The purpose of this post
  + The guessing game setup
* Defining measures of central tendency
  + The mode
  + The mean
  + The median
* Deriving measures of central tendency as solutions to variational problems
  + Cost functions
  + Inequality penalty
  + Absolute difference penalty
  + Squared difference penalty
  + L^p metrics
  + Max difference penalty

Outline version 2.1

* Introduction
  + The purpose of this blog is to explore how we can reason about what we know to make guesses about what we don’t.
  + We are going to “dip our toes” into this subject by exploring one of the most commonly experienced concepts in probability theory and statistics: measures of central tendency.
  + We are going to revisit the common definitions of measures such as the mean, median, and mode; and attempt to build a stronger intuition of what they are and why they are useful in our goal of making better guesses.
  + Before we begin, we are going to frame our discussion in terms of trying to win a particular guessing game.
* The Guessing Game
  + Suppose we are asked to participate in a guessing game.
  + Specifically, we are asked to predict what number will be printed on a ball drawn from a bin containing 100 numbered balls.
  + The next ball will be drawn “randomly”, that is to say in a manner that makes any ball equally likely to be drawn.
  + Note however that this does not mean that the numbers printed on the balls are equally likely i.e. some numbers appear on more balls than others.
  + Let’s make this setup concrete by creating a bin of balls in R.
  + print(x)
  + Here we can see the balls in our bin. Some numbers occur more than others.
  + Perhaps a better way of viewing this is by plotting how frequently we see each number in our bin.
  + This frequency is closely related to the concept of probability, and in some schools of thought probability is defined as this frequency.
  + We can see this relationship by normalizing the counts to be between 0 and 1 by dividing by the counts by the total number of balls.
  + We can now say that these normalized counts aka relative frequencies represent the probability that the given number will be drawn next.
  + So, what number should we guess?
* Defining Measures of Central Tendency
  + A measure of central tendency can be defined as a value that represents a typical value in a collection.
  + The three most commonly encountered measures are the mean, the median, and the mode.
  + We are going to define each measure in turn. Some of this will likely be review, but we hope to bring to light some properties that relate measures of central tendency to making reasonable guesses.
  + A quick primer on types of data:
    - Nominal data:
      * observations take values that are names (nominal comes from the Latin for name), or labels more generally.
      * For instance, we could observe a ball is “green”; a human is “female”; or a product has a barcode “1099231”. In all three cases, the value relates the object of our observation to the name of some category or class.
      * If numbers are used, we say they are nominal if the numbers don’t have meaning e.g., they do not communicate an order, and differences or ratios are meaningless.
    - Ordinal data:
      * Observations take values that reflect some sort of “order”.
      * For instance, we could observe a runner finishing “second” in a race; or a person responding that they “strongly agree” with a survey question.
      * If numbers are used, we say that they are ordinal if the order matters i.e. 2 < 3, but the difference between numbers is meaningless, so we can’t say that 3 - 2 = 4 - 3
    - Interval data:
      * Observations take values that are numbers where both order and differences are preserved. The limitation of interval data is that there is no true zero value (think of temperature on the Celsius or Fahrenheit scale). This means we cannot compare the relative magnitudes of measurements (ratio), just the differences.
    - Ratio data:
      * Observations take values that are non-negative numbers where order, differences, and relative magnitudes are preserved. The full range of comparison is available for ratio data.
  + The mean:
    - The mean is usually what people refer to as the “average” value.
    - Simply, the mean is the sum of the value of each observation divided by the total number of observations. This is often referred to as the “sample mean” when applied to a finite set of observations.
    - More generally, the mean is referred to as an expectation when we consider it as the sum of all possible values weighted by their probability of occurrence.
    - This definition provides a lot more generality and directly relates this measure of central tendency to the probability of outcomes.
    - It should be noted that the sample mean is only defined on interval and ratio data. Ordinal and nominal data do not have a mean value.
  + The median:
    - The median is the second most commonly encountered measure of central tendency.
    - Simply, if you were to line up all of your observations by their value, the median would be the value in the middle.
    - When there is an even number of observations, we take often define the median as the average of the two middle values, but technically the median is any number between these two values.
    - A more generally understanding of the median occurs when we take a different vantage point of our probability distribution, specifically when we look at the cumulative distribution function.
    - Here we can see that the median value is the value that we have a 50% chance that an outcome will be less than it and a 50% chance that an outcome will be greater than it.
    - The median can be defined for ordinal, interval, and ratio data.
  + The mode:
    - The mode is simply the most frequently occurring value in a collection of data.
    - More generally, the mode is the value that has the highest probability of occurring.
    - There may be multiple modes in a distribution (this is referred to as being multimodal), and there may be no/infinite modes depending on how we wish to define the mode.
    - There is a nifty interpretation of the mode from information theory: the mode is the value that we are “least surprised” to see when it appears.
    - The mode can be defined for nominal, ordinal, interval, and ratio data.
  + All three of these measures of central tendency all seem reasonable, but which one do we choose as our guess?
* Deriving Measures of Central Tendency
  + In order to answer the question “do we choose the mean, median, or mode as our best guess?”, we can put our question back into the context of our guessing game.
  + Our game is missing a critical feature found in most games: a method of scoring.
  + In mathematics (specifically optimization) we use cost functions (synonymously loss functions, alternatively utility functions, and more generally objective functions) as a means of scoring possible solutions to problems (in our case possible guesses).
  + With cost functions, points are bad (like in golf) and we want to minimize the number of points we receive during the game.
  + In order to determine what our best guess is, we need to know how we are scoring the game i.e. how much a wrong answer will cost us.
  + We are going to consider a number of scoring systems aka cost functions, and how they relate to the measures of central tendency we have already discussed.
  + Inequality penalty:
    - The first cost function we are going to consider is an inequality penalty.
    - The idea is simple: we get 1 point when we guess wrong, and we get 0 points when we are exactly right.
  + Absolute difference penalty:
  + Squared difference penalty:
  + L^p metrics
  + Maximum difference penalty:
* Aftermath