Data Literacy

William John Holden

Introduction

Parameters and statistics

Statistics are the foundation of most data mining, machine learning (ML), and artificial intelligence (AI) methods today. A *statistic* is an estimate of a *parameter*, which is a characteristic of an entire *population*. Statistics are calculated from taking *samples* (subsets) from the population.

For example, suppose we wanted to find the height of the tallest mountain in the world. We might sample $n = 100$ mountains at random from an almanac. Suppose the tallest mountain in our almanac is Mount Fuji, the tallest mountain in Japan, which is 3776 meters tall. We can safely conclude that the tallest mountain is *at least* 3776 meters tall.

Our estimate is unfortunately quite low. Mount Everest in Nepal, the *highest* mountain in the world, stands 8849 meters above sea level. Mauna Kea in Hawai'i, the *tallest* mountain in the world, stands 4207 meters above sea level and another 6004 meters below. Our estimates of population parameters, *statistics*, generally improve with larger sample sizes, and many statistical methods provide a *margin of error* quantifying sampling error.

One might use statistics to create a *model* to explain a population, based upon sampling data. Models can be useful both for describing the population and also for forming predictions.

Nominal, ordinal, interval, and ratio variables

There are several classes of data that a variable might fit into. *Nominal* data is simply names or categories, with no concept of order or distance. A movie might be animated or live-action: these are simple categories or order. The name of the nation where the movie was filmed is another example of nominal data. Simple yes and no categories are also nominal data, such as whether a film does or does not have a sequel.

Ordinal data has ordering but not distance. Ordinal data might be represented as ordered categories or as numerals, though these numerals do not provide

meaningful addition and subtraction. The ratings of a film (G, PG, PG-13, R, and so on) form a ranking, but addition is meaningless (does G + PG-13 = R?) and our concept of distance is weak at best. Another example of ordinal might be the rankings the films receive at an awards ceremony, where one film is the winner and another is the runner-up.

Interval data is numerical data with a concept of distance but not multiplication. The year when a film was produced is an example of interval data. If two films were produced in 2000 and 2010, then it makes sense to say one was made ten years later, but we would not say that the latter film is 2010/2000 = 1.005 times the first.

Ratio data is numerical data with both distance and multiplication. The gross earnings of a film is an example of ratio data. If the 2000 film earned one million dollars and the second earned two million dollars, then it makes sense to say the second film earned double the first.

Name	Operations	Type
Nominal	$=, \neq$	Categories
Ordinal	$<, >$	Ordered categories
Interval	$+, -$	Numbers with distance
Ratio	$\times, \div$	Numbers with meaningful zero

Interval data might be initially confusing to distinguish from ratio data. One indication is the absence of a meaningful zero. Does zero degrees Celsius or Fahrenheit mean the absence of temperature? No, these measurements are simply points along a scale. Twenty degrees Celsius is not "twice" ten degrees Celsius; multiplication is not defined on interval data.

Grid coordinates might be another example of interval data. One can calculate the distance between two grid coordinates, but we would not say that coordinate 1111 is "half" of coordinate 2222.

Data might be represented in numerical formats when some operations do not make sense. Suppose a political scientist encoded voter's political party as "1", "2", "3", and "4". Is "2" an intermediate value between "1" and "3", or are these actually nominal data where the only arithmetic operations are = and $\neq$? AI methods sometimes make incorrect assumptions about data that domain experts can easily prevent.

Discretization

Measurements with arbitrarily many decimal digits of precision are *continuous*, whereas measurements with finite steps in between (including categories) are *discrete*. For example, when driving along a road, the house numbers (150 2nd Street, 152 2nd Street, 154 2nd Street…) are discrete; there is no intermediate

value between 150 and 151. On the other hand, the grid coordinates associated with each address are continuous; one could (theoretically) specify grid coordinates to the nanometer.

It can be useful to combine continuous measurements into discrete categories. An example might be one's birth date and birth year. No one knows their birth *instant* with subsecond precision. Rather, the year, year and month, or year, month, and day are almost almost always enough information. We even combine years into groups when discussing generations and peer groups. Combining a range of birth years into generational categories is an example of *discretization*.

Missing values

In practice, data sets are often missing values. Different programming languages have substantially different syntax and semantics for representing and handling missing values.

As a small exercise, open Microsoft Excel and enter the values 1, 2, 3, and 5 into cells A1, A2, A3, and A5. Leave cell A4 blank. In cell A6, enter the formula `=PRODUCT(A1:A5)`. The result is $30 = 1 \cdot 2 \cdot 3 \cdot 5$. Excel did *not* treat the missing value as a zero.

Now change cell A4 to `=NA()`. `NA` means "value not available", an explicit indication that a value is not given. The product in cell A6 should update to `#N/A`, which explicitly tells us that there is a problem in the calculation.

Now change cell A4 to `=1/0`. Both cells A4 and A6 should both say `#DIV/0!`, a fault telling us that a division by zero has made further calculation impossible.

Error values propagate from source data through intermediate calculations to final results. If we enter a formula into A7 referencing A6, such as `=SQRT(A6)`, then we will find the same faults in A7 that we see in A6.

Structured Query Language (SQL) databases use the symbol `NULL` to denote missing values. One might build the database *schema* (the structure of the database) to explicitly forbid `NULL` values. For example, `CREATE TABLE Race (Name TEXT NOT NULL, Time INTEGER NOT NULL)` creates a table of run times where both the name and the time must be specified.

Many programming languages support a `NaN` ("not a number") value in error conditions. One might encounter `NaN` when dividing by zero, subtracting infinities, and parsing non-numeric words as numbers. Comparisons with `NaN` can be confusing, such as `NaN == NaN` returning *false*.

Some programming languages will automatically *initialize* variables with some zero value. Other languages give some `Undefined` value to uninitialized variables. Still other languages raise an error if no explicit value is assigned to a variable.

Strong/weak and static/dynamic typing

Columns and rows

Features and individuals

Tables, lists, and data frames

Vectors and matrices

Box plot, scatter plot, bar plot, and histogram

Linear and logarithmic scales

Functions

Discussion prompts

1. Who owns knowledge management?
2. What are good and bad uses for spreadsheets?
3. What is reproducibility and why would this be important for scientific inquiry?
4. Why is a pie chart not recommended?

Practical exercises

1. Given a dataset, plot the data and explain why this plot technique is appropriate.
2. Given a noisy and poorly structured dataset, propose a method of restructuring the data.
3. Discretize the values of a dataset and explain the reasoning.
4. Be creative and construct intentionally misleading plots that deliberately distort information presented.

Data Operations

Database schema

Forms and input validation

Select, project, and join

Filter, map, and reduce

Grouping and aggregation

Vectorized functions

Concurrency

Consistency, availability, and partition-tolerance (CAP) theorem

Discussion prompts

How does the CAP theorem impact intelligence and fires in relation to the command and control (C2) warfighting function (WfF)?

Where should unclassified data be stored and processed?

What are some methods to prevent conflicts among concurrent writes in a shared database?

What could possibly go wrong when altering database schema?

Practical exercises

Create a custom list in SharePoint that provides multiple views showing grouped and aggregated values.

Given a noisy dataset, identify problems in each column that could influence inclusion and exclusion criteria.

Implement filter and map in terms of reduce using a programming language which provides reduce.

Define an “embarrassingly parallel” problem and provide both examples and counterexamples.

Measures of Central Tendency

Mode

Median

Arithmetic Mean

The four moments: mean, variance (and standard deviation), skewness, and kurtosis

Exponential moving averages (EMA)

Covariance

Outliers

Unbalanced data sets

Discussion prompts

Is four a lot?

First battalion has an average ACFT score of 482 while second battalion has an average ACFT score of 491. Which is better?

What do we do when statistics show us something that contradicts our values? For example, suppose we discover that Soldiers of a specific demographic have much lower promotion rates than their peers.

Is it more important for an organization to think about variance or the 99th

percentile?

Given a sample set [Equation], what is the estimate of the mean ([Equation]), and what is the sample variance?

Practical exercises

Calculate the influence that outliers have on different-sized datasets that contain outliers.

Calculate the exponential moving average in a small dataset.

Given a dataset and experimental result, identify problems caused by analyzing categorical data represented in a numeric form.

Given multiple datasets with identical mean and standard deviation, use kurtosis to identify the dataset with more outliers.

Design or implement an algorithm to incrementally calculate standard deviation, where the estimate of the sample standard deviation is updated with each additional value.

Linear Models

Sum of squared errors

Linear regression

Polynomial models

Pearson correlation coefficient

Prediction

Overfitting

Correlation and causation

Discussion prompts

What are confounding factors and how can we identify them?

When is it useful to generate a linear model with no previous understanding of the data?

Some data mining techniques fail if the data set contains colinear columns. How might one identify these columns?

Practical exercises

Design a novel linear model, using any programming language, that considers time when fitting [Equation] values and creates a biased model (like an EMA).

Given linear models and their associated [Equation] values among many different columns in a data set, identify the strongest and weakest models, then use plots to support these conclusions.

Using the Goal Seek feature of Microsoft Excel, estimate the mean of a column by minimizing the sum of squared errors. Compare this value to the arithmetic mean.

Use more than one tool (such as both Excel and R) to find the line of best fit on a single dataset. Compare and contrast the linear regression generated by each program.

Hypothesis Testing

Combinatorics

The Binomial Distribution

The Normal Distribution

The Central Limit Theorem

Null hypotheses ([Equation]), [Equation]-values, and [Equation]-values

Probability density function (PDF) and cumulative distribution function (CDF)

Student's [Equation]-test

Pearson's [Equation]-test

Analysis of Variance (ANOVA)

Discussion prompts

How does the "curse of combinatorics" create effectively infinite event spaces?

Suppose a daycare has 1000 toys in the toybox. Each time a child takes a toy from the toybox, a worker records the toy and the (seemingly independent) result of a fair coin toss. After each toy has been pulled ten times, the worker discovers that the coin always landed on heads for the toy shark. How surprising is this outcome?

A study shows [Equation]. What does this result mean?

Describe a workplace situation where the Central Limit Theorem applies.

Practical exercise

One group generates a random data set that should not fit a normal distribution. Another group takes samples and applies the Central Limit Theorem to estimate the population mean.

Given a multivariant dataset, use ANOVA to identify the strongest linear relationship between a dependent variable and many independent variables.

An obscure Filipino superstition claims that the gender of the firstborn child predicts the first parent who will die (e.g., if the eldest child is a son, then the father will pass away before the mother). Create a small data set by surveying the class, then use the [Equation]-test to analyze the strength of the superstition.

Use a CDF/PDF implementation to convert a uniform random number generator (RNG) into a normal distribution.

Supervised Learning

Learning from data

Test-train split

Confusion tables and model accuracy

Decision trees

Artificial neural networks (ANN)

Activation functions

Backpropagation

Discussion prompts

Why might a voice recognition model more reliably understand one accent than another in the same language? How might the model be improved?

What happens when a model is trained on information that the model itself produced?

Unlike linear models, neural networks are often initialized with random values. As a result, training two models on the same data might not lead to identical results. What are the ethical implications of having different results in different models?

Practical exercises

In any programming language, train an ANN to recognize handwritten characters. Use a test-train split to evaluate the accuracy of the model.

Create a classification model for an extremely unbalanced data set, then assess the accuracy of the model.

Create a classification tree by hand to identify Russian military vehicles.

As a group, create an "ensemble" of different models that vote on the outcome of a classification task.

Unsupervised Learning

Data mining

Principal component analysis (PCA)

Hierarchical clustering

[Equation]-Nearest Neighbors (kNN)

[Equation]-Means Clustering

Constraint solvers

Embeddings

Word2Vec

Discussion prompts

Suppose a dataset contains a sequential numeric identifier for each row. An unsupervised learning method unexpectedly uses this feature to predict an outcome with good accuracy. What does this mean?

Supposes an embeddings model produces "queen" from "king – man + woman", but unexpectedly produces "king" from "queen – English + Turkish – Turkish + English". Why might this occur?

Practical exercises

Compile raw ACFT data for the class, perform PCA, and attempt to interpret the first and second components.

Using the same ACFT data, use cluster analysis to attempt to predict Army component (RA, AR, NG). Use the diagonal of the resulting confusion table to assess model accuracy.

Model a scheduling problem using a constraint solver.

Graph Theory

Vertices and edges; [Equation]

Directed and undirected graphs

Directed acyclic graphs (DAG) and topological sorting

Weighted graphs

Breadth-first search (BFS) and depth-first search (DFS)

Dijkstra's algorithm

Computational complexity and Big-[Equation]

Graph databases

Power Law Distribution

Discussion prompts

A graph can be represented with an adjacency list or a matrix. What are the advantages and disadvantages of each approach?

What algorithm can be used to solve the "seven ways to Kevin Bacon" problem?

Is a Gantt chart a graph? How can one find the critical path of a project if represented as a graph?

Which is bigger, [Equation] or [Equation]?

Practical exercises

Compare two different heuristic functions in a provided A* informed search implementation on the 8-piece puzzle problem.

Convert currency exchange rates from multiplication to addition using a logarithm, then prove that infinite arbitration is impossible given a set of exchange rates and Bellman-Ford implementation.

Define a topological sorting and relate it to a workplace problem.

Define the Traveling Salesman Problem (TSP) and explain the computational difficulty of this problem.

Determine the minimum paving needed to fully connect a tent complex using a list of coordinates and a Prim or Kruskal implementation.

Simulate an infection model in a dense social graph where edge weights represent probability of infection.