Interpreting Problems and Results

This strand focuses on justification and explanation of reasoning when making inferences, claims, or suggestions from data within the context and processes of the dataset collection and analysis.

An important component of interpreting results is understanding the relationship between questions, problems and datasets. Formulating a strong question or identifying a problem that can be addressed with data affects the opportunities for interpretation and results from the data. Additionally, the applicability of inferences and claims that are made are constrained by the sample, population, and context of the data.

Substrand D1

Making and Justifying Claims

As all data contains variability, it is important to use probabilistic thinking and language when making claims from data. This requires paying attention not only to patterns and comparisons within and across variables but also such things as expected and prior values, sample sizes, and significance.

Concept D.1.1

Probabilistic language

When communicating with others, employ both plain-language and clear vocabulary to regularly describe degrees of uncertainty, both formally and informally as a thinking habit.

21st-century skills

Durable skills

Clearly state a result or finding, along with the degree of certainty, using two or more advanced statistical methods (e.g., probability distributions, t-tests, z-tests, or bootstrapping/simulation), while justifying the conclusions with evidence (e.g., dataset or source characteristics, similar findings in alternative data) quality indicators like dataset characteristics, source reliability, and corroborating findings from alternative data.

Concept D.1.2

Priors and updates

When encountering new data, integrate probabilistic thinking into everyday situations by explicating prior assumptions and the impact of new data / evidence on those assumptions.

21st-century skills

Durable skills

Media literacy and digital citizenship

Explain Bayes Theorem in formal conditional probability statements: P(A|B) = (P(A) * P(B|A)) / P(B), where A is the event in question and B is the event of new evidence related to A.

Apply Bayes Theorem to an example result in an academic research finding or discussion.

Concept D.1.4

Explaining significance

Clearly describe the basic logic of statistical significance to others, differentiating between significance, the size of an effect, and the statistical power of an analysis. Recognize what statistical significance can reveal and cannot reveal about a phenomenon.

21st-century skills

Media literacy and digital citizenship

Identify examples of p-value misuse in the media or academic research.

Describe a p-value to without using the language of the "null hypothesis" or "alternative hypothesis."

Concept D.1.5

Sampling and simulation

Comfortably identify the purpose of sampling and simulation for making arguments about data, and employ techniques using software to differentiate a real-data result from random chance or “happenstance.”

Execute and correctly interpret the margin of error, confidence interval, and standard deviation in a data analysis software for a given summary statistic.

Describe the relationship between the margin of error, confidence intervals, and standard deviation, in both words and in their formal mathematical definitions.

Concept D.1.6

Correlation versus causation

21st-century skills

Media literacy and digital citizenship

Comfortably separate correlation from causation in a wide variety of situations, building a “first-reaction” thinking habit over time.

Justify a causal relationship in a multivariable dataset with real-world data, including additional datasets gathered from outside sources and connect the analysis to existing research literature.

Explain why a chosen analysis method effectively isolates an effect.

Concept D.1.7

Randomization

When identifying a potential cause of a phenomenon, clearly describe the usefulness of randomization for constructing an argument with data.

Implement randomization using a random seed in a simulation technique using a computer-based analysis tool to compare sampling techniques (e.g., sampling with replacement or without replacement).

Concept D.1.8

Multi-variable decision-making

21st-century skills

Durable skills

Clearly describe how to leverage additional variables or additional outside data to make a logical argument, and identify potential risks of overdoing it.

Use computer software to incorporate categorical variables into a linear regression model.

Analyze and interpret the regression coefficients to understand the effect of the categories on the model.

Create an “ideal” multi-variable model for real-world data in a computer-based software that explains as much variance as possible, without overfitting a model. Justify how you have found the “ideal” model by comparing R^2, covariance, and the number of variables chosen in relation to their real-world context.

Substrand D2

Problem Identification and Question Formation

Formulating a question or identifying a problem that can be addressed with data affects the opportunities for interpretation and results from the data. The ability to make and justify strong claims relies on identifying questions that are testable and can be answered with data. Additionally, identifying the uncertainty or limitations within the problem space is an important component of formulating conclusions

Concept D.2.2

Iteration, validation, and multiple explanations

21st-century skills

Durable skills

Regularly practice identifying alternative explanations for a result from data, both for interim steps and post-analysis conclusions.

Document analysis steps and errors while implementing validation checks in the software for data wrangling.

Execute an alternative analysis plan to validate a significantly different result from the initial method.

Substrand D3

Generalization

Though there is often an instinct to use data to make large generalized claims, the applicability of inferences and claims that are made are constrained by the sample, population, and context of the data.

Concept D.3.2

Sample versus population

Given a dataset, identify constraints and opportunities for what can be logically inferred about a broader population.

21st-century skills

AI literacy

Identify machine learning methods such as supervised, unsupervised, and reinforcement learning, and discuss the pros and cons of each when data on the entire population or a very detailed sample with many variables is available.

Concept D.3.3

Sample size

When full information is hidden or inaccessible, recognize the logical relationship between a sufficient number of chances and a sufficiently large sample to reasonably represent something.

21st-century skills

Media literacy and digital citizenship

Make a formal Power Analysis by identifying a sufficient sample size for a real-world data exploration. Students should mathematically isolate “n” in a t-test or z-test, and estimate Power with a software tool.

Concept D.3.4

Simple bias

When information is completely hidden or unavailable, be aware of possible underlying issues in the sample and apply strategies to identify and address them.

21st-century skills

Media literacy and digital citizenship

Estimate bias by interpreting and applying the formula for a biased estimator.

Strand E

Tools and Techniques

Visualization and Communication

This strand focuses on how to communicate about data through the creation and examination of visualizations.

Visualizations are a vital component of the sensemaking process when working with data. Being able to communicate with and about data using visualizations that are clear and tailored to a purpose and audience are an important step for creating action and impact through data. Also important are skills and habits for how to read, interpret, and critique other's data communication, paying attention to context, audience and purpose.

Substrand E1

Representations and Dynamic Visualizations

The creation and interpretation of graphic and interactive visualizations are vital components of the sensemaking process when working with data. Working with data visualizations requires an understanding of conventional components and best practices along with graphical literacy and representational fluency.

Concept E.1.1

Sense-making with visualizations

Practice creating visualizations to summarize many things at once, relationships between things in one place, or exceedingly complex ideas in one place. Recognize that visuals can be more efficient or compelling than other forms of communication.

Demonstrate presentation skills to fully communicate depth and breadth of a visualization to an audience.

Present both
1) basic visual summaries of the data
2) additional visualizations that "go deeper" into the story the data is telling, and relationships discovered within the data visualization
e.g., new relationships within subsets, significant outliers, complex or overlapping control variables

Concept E.1.3

21st-century skills

Durable skills

Clear design for user interpretation

Identify conventional components and best practices of data visualization from a user-centered or audience perspective.

Apply design principles such as balance, emphasis, and simplicity to make visualizations clear and engaging.

Understanding the basics of interactive visualizations (e.g., tooltips, zooming) and their advantages in data exploration.

Concept E.1.4

21st-century skills

Media literacy and digital citizenship

Graphical literacy

Comfortably read graphs with accuracy and make sense of data visualizations by answering questions about how the data is represented with precision.

Describe the potential relationships (or lack thereof) represented in scatterplots (including linear, exponential, logarithmic, polynomial, and piecewise) and debate which function is the best representation for the shape and context.

Visualize confidence intervals or margins of error using error bars with computer-based software.

Visualize margins of error of a continuous variable using error bands with a computer-based software.

21st-century skills

Media literacy and digital citizenship

Concept E.1.5

Representational fluency

Identify how layout (ordering, scale, and axes) choices increase clarity or potentially mislead an audience.

Compare and/or contrast 2D and 3D bar graphs and pie charges and identify how unnecessary use of three dimensions obfuscates the relative frequencies and/or proportions of the data.

Compare and/or contrast varying bin sizes to demonstrate how different degrees of granularity in a histogram or other visualization type can lead to different interpretations.

Substrand E2

Data Storytelling

Being able to communicate with and about data using visualizations connected to a narrative is an important step for creating action and impact through data. Understanding the audience for the narrative is vital to clear communication.

Concept E.2.2

Write data stories

Structure effective stories about data when complex jargon and technical ideas are involved.

21st-century skills

Durable skills

Use complex visualizations like multivariable graphs, scatter plots, heat maps, or interactive dashboards to present data clearly. Then, develop a research paper or presentation to explain the background, methodology, and context of the data, using visualizations to provide evidence of their findings and conclusions.

Substrand E3

Acting on Data to Benefit Society

One of the ultimate goals of working with data is applying interpretation and conclusions to real-world problems and scenarios in order to engage in civic practice and enact positive change on the world.

Concept E.3.3

Civic data practices

Engage in civic practice and dispositions through recognition of the role data plays in civic society.

Pick a local issue of student interest and draft a Letter to the Editor (LTE) to a local news outlet or to a local politician based on conclusions from public-access datasets.