Center & spread

Joseph Roith

2019-04-22

Activities

- Variables and their summary intervals
- The standard deviation and the summary interval
- Shapes of distributions
- The normal distribution

Learning objectives

1. Frame the prediction of an unknown quantity using an *interval*.
2. Choose the prediction interval endpoints to convey a range of likely outcomes.
3. Understand that predictions are more likely to be right when the interval is constructed to be long.
4. Understand, in contrast to (3), that predictions are more likely to be informative when the interval is short.
5. Master the convention that is widely used for accomplishing both (3) and (4), e.g. the use of 95% coverage as a sensible evaluation of the trade-off. Identify 95% as a "level" selected specifically to have a reasonable balance between (3) and (4).
6. Be able to work with two formats for describing an interval: "A to B" and "C ± D"
7. Use intervals as a way to identify outliers and data anomalies.

Additional sections

-
- Instructor orientation
- Role in statistical practice
- Classroom discussion
- Tips for an active classroom
- Student pre-requisites
- Pitfalls

Orientation for instructors

Early in most introductory statistics courses, students are taught ways to describe quantitative variables. As regards numerical descriptions, the mean and standard deviation are "normal." (Pun intended.) The median and interquartile range offer an alternative when the variable is "skew." The means and/or median are described as "representative" or "typical" values. The standard deviation describes how far "typical" values are from the mean, that is, how much "typical" values are spread out.

It's laudable that instructors look for ways to enliven or renew such topics, and to seek data and contexts that will motivate students to appreciate the importance of ideas such as "representative," "typical,", and "spread." RESOURCES FOR THIS, e.g. bag of tricks

In these StatPREP lesson notes, we're going to take a less trodden path, recognizing there are already plenty of guides to the conventional route. So rather than orient the lesson around "center and spread," we're going to treat as the core ideas describing variation and framing predictions.

Role in statistical practice

- Variation
  - understand: it = the variation. How it behaves, how it reacts under different criteria, situation. how do we expect from one observation to another, from one person to another, how a quantity varies.
  - We want to quantify it to
    * describe typical
    * understand how future observations behave
- Part of exploratory data analysis
- Identifying odd values

Conceptual pitfalls

What does "representative" mean? Is "typical" a single value?
Standard deviation:

- Daunting name which unnecessarily inclines the student mind to relate the quantity to negative social constructs such as "deviancy," "abnormal",
- Standard pedagogy understandably makes standard deviation a derivative of center. Center comes first, standard deviation follows.
- The calculation brings in seemingly arbitrary choices: squares, $n - 1$. Even many instructors find $n - 1$ mysterious.
- There's no room for student exploration: it's a pat quantity.

Student pre-requisites

- Graph reading:

    - the presentation of individual values of a quantitative variable as a point in a graph.
    - identify the most common values using the density of such points.
    - translate density into a length for graphics purposes (e.g. histogram, violin plot)
    - read off from a point plot reasonable answers to these questions:

        * What's a high value?
        * What's a low value?
        * What's a middle value?
        * What is the range of the data?
        * What values are just completely out of range?
        * Are there individual values in the data that are so far from the bulk of other values as to reasonably be considered out of range for practical purposes?

    - familiarity with plot types for presenting a quantitative variable against a categorical variable

- Data:

    - correctly distinguish between quantitative versus categorical variables
    - understand what it means to describe the unit of observation in a data frame
    - have vocabulary to distinguish between a single observed row of a data frame ("case", "row", sometimes "person") and the collected set of many rows.

Classroom discussion

- the prediction game

- best placement of intervals

- should intervals be short or long?

- quantifying big, small, "slightly" different, unsurprising difference, surprising difference. Defining big in terms of a the individual variables, e.g. a couple of standard deviations.

- how an explanatory variable can make prediction better

- prediction as a framework for talking about difference

- Why do we leave some cases out of the interval?

    - outliers, dilutes what we have to say

- Equivalence of top-to-bottom and middle +- width

- Train eye to say what range covers almost everything. Then, give a meaning-ful standard for "almost everything."

- How well does a given number represent the center of the distribution?

  - most people are that height: no!
  - normal people are that height: but what's "normal"
  - most people are "close" to that height?

    * What does "close" mean?
    * What does "most" mean. "representative": majority,

- How do we answer questions like "about the same" or "very different". Or "two random people will be about this close together"

- Let's play a little game. Tell me what the temperature is outside. You pay $5 to play. You'll win $100 if you guess the exact right temperature. Would you play that game?

  Change the game. You give me a range of temperatures. You'll get $1,000,000 if the actual temperature is in that range.

  What if they had to pay you $1 for every 0.1 degree in the range.

Creating an active classroom

Let students discuss among themselves:

- how to evaluate the tradeoff between stating narrow prediction intervals and stating intervals that are likely to contain the eventual value when it is known.
- comment on the choice of 95% level and the plusses and minuses of that compared to a 50% or other level.

Pushing the envelope/advancing the field

- a proper statistical prediction is not simply a number, but expresses both belief and uncertainty.
- standard deviations are one of the ways of describing spread, but there are many others, such as 95% summary intervals.

Author info