Outliers

Contrary to many overzealous ethics sticklers, discarding data that do not fit a trend is part of normal scientific practice in many cases. The reason is simple: the universe is complicated. The mechanisms producing your data are not always the mechanisms you want to study. In astronomy, this can take the form of dead pixels or hair and dust on the camera.

Dead pixels wipe out a column of a telescope image showing oversaturated white stars on a pixellated blue background.

Dead pixels wipe out a column of a telescope image. Image from here.

In cell biology, a model might assume RNA splicing reactions represent "on" and "off" steady states, when in fact they are sometimes from a transient intermediate state.

A figure from the RNA velocity paper showing robust estimation of equilibrium splicing ratios based on the extreme quantiles of the phase plot. Figure shows spliced (x axis) and unspliced (y axis) RNA levels, with different color lines for the true slope, the regular fit, and the quantile fit.

Robust estimation of RNA splicing dynamics from supplemental note 2 of this awesome paper.

In economics, the money supply followed predictable yearly bumps and whorls – until a raging coronavirus pandemic shut down all the shady businesses that mobsters had previously used to launder wads of cash.

Time-series of Canadian money supply shows yearly bumps and eddies and a CAD 20 billion takeoff in April to August 2020.

Canadian money supply during the early days of the pandemic, from JP Koning.

In today's session, we will illustrate the principle that sometimes, *an estimator must not respond to outliers.*

Learning objectives

- Experience the effect of outliers on the mean, median, and mode of a symmetric Binomial distribution.

Exercises

1. Look up the definitions of the mean, median, and mode (Khan Academy explains nicely here). How should they each react to one extreme outlier?
2. Generate a random sample of size 1,000 from a Binomial distribution with 100 trials and p=0.5. Use the function `rbinom`. Plot a histogram of the data using `hist`.

3. Compute the mean, median, and mode of the sample.
4. Find a classmate and ask them to add one outlier to your dataset. Be creative.
5. Add the outliers to your sample. Compute the mean, median, and mode.

More exercises

The data for the rest of the session comes from some research I've been working on recently. It's from a simulation of a small network of 18 genes that interact with one another. The DNA of each gene is transcribed into RNA, and the RNA is translated into proteins. The rate of transcription of RNA depends on certain protein products from the other genes. The rate of RNA decay is simpler. It is proportional to the amount of RNA present. The production rate minus the decay rate is the total rate of change of the RNA, also called the RNA velocity. The simulation reveals the amount of RNA and the RNA velocity, but without separating production and decay.

```
rna_velocity = rna_production - rna_decay
rna_production = ???
rna_decay = d*rna_quantity
rna_velocity = ??? - d*rna_quantity
```

For my research, I needed the production and decay rates separately, not just the total velocity. Let's look at the data to see if there's any way of teasing them apart. Here's a typical gene from this dataset. Each dot represents the values at the final time-point of a separate run of the simulation. There is some randomness purposefully included. We'll look at three quantities: the time the simulation was terminated; the RNA level of gene 7 at that time; and the rate of change of that RNA level.

A time-series of RNA levels begins at near-zero expression with little variability and rapidly grows to high expression with high variability around time-point 150. A bivariate display of RNA concentration and RNA velocity shows two parallel lines with negative slope, one at low RNA levels and the other at high RNA levels, plus some wild outliers, which are mostly in-between.

A time-series of RNA levels begins at near-zero expression with little variability, and around time-point 150, it rapidly grows to high expression with high variability. A display of RNA concentration and RNA velocity shows two parallel lines with negative slope, one at low RNA levels and the other at high RNA levels, plus some wild outliers, which are mostly in-between.

I chose each chart for a rhetorical reason.

- The first plot gives us important context about this dataset: gene 7 has two steady states, "on" and "off", with a rapid transition in between.

- The second plot tells us that within each steady state, the velocity decreases in an nearly-exact linear relationship with the RNA concentration. This would happen if the production rate was constant and the decay rate dominated the trend. It could also happen if the production rate was exactly linear too, but based on what I know about the dataset, that's not true. Let's assume the slope of the line is the decay rate `d`.

That means the decay rate we need is the slope of the two parallel lines dominating the second plot. In this exercise, you will try to estimate the slope in a way that ignores outliers. Email your R code and report to Eric as `LAST_FIRST_outliers.R` and `LAST_FIRST_outliers.docx`.

1. Download the data and the starter script. The columns for genes 7 and its velocity are `x_g7` and `velocity_x_g7`. Reproduce the two plots above.
2. Add a least-squares line of best fit to the second plot. This is analogous to a sample mean. You can using the R function `lm` to find the slope and intercept, and you can use the `abline` function to add a line with the same slope and intercept to the plot. Does this seem to capture the trend we are looking for?
3. There is another type of line of best fit analogous to the median called *quantile regression*. Add a quantile regression line of best fit. You can use the `quantreg` R package, but you probably will need to install it first with `install.packages("quantreg")`. Does this seem to capture the trend we are looking for?
4. How else could you capture the trend in this dataset? Get creative and come up with a better solution. If you're stuck, ask for a hint.
5. Wrap your code in a single function so you can test least-squares, quantile regression, and your solution easily on another gene. Produce the same plots for genes 1, 5, 16, and 18. Does your solution still work? If not, can you modify it to be more general?