JSM 2021 STUDENT PAPER COMPETITION (ASA SECTIONS ON STATISTICAL COMPUTING AND STATISTICAL GRAPHICS)

Perception of exponentially increasing data displayed on a log scale

Emily A. Robinson[a], Reka Howard[a], Susan VanderPlas[a]

[a]Department of Statistics, University of Nebraska - Lincoln,

ARTICLE HISTORY
Compiled November 10, 2020

ABSTRACT
Log scales are often used to display data over several orders of magnitude within one graph. During the COVID pandemic, we've seen both the benefits and the pitfalls of using log scales to display data. This paper aims to...

KEYWORDS
Exponential; Log; Visual Inference; Perception

1. Introduction and Background

Graphics are a useful tool for displaying and communicating information. Researchers include graphics to communicate their results in scientific publications and news sources rely on graphics to convey news stories to the public. During the onset of the novel coronavirus - covid19 - pandemic, we saw an influx of dashboards being developed to display case counts, transmission rates, and outbreak regions (Charlotte 2020). As a result, people began subscribing to news sources involved in graphically tracking the coronavirus (example John Burn-Murdoch Financial Times - CITE THIS) and gaining more exposure to the use of graphics. Many of these graphics helped guide decision makers to implement policies such as shut-downs or mandated mask wearing. Better software has meant easier and more flexible drawing, consistent themes, and higher standards.You need to source this somehow, please. Otherwise it seems disconnected from your main argument. Consequentially, we must develop a set of principles to help us actively choose which of many possible graphics to draw (Unwin 2020).You might need to lead into this a bit more - connect the purposes we use these graphics for to the fact that different choices serve those purposes in different ways.

When faced with data which spans several orders of magnitude, we must decide whether to show the data on its original scale (compressing the smaller magnitudes into relatively little area) or to transform the scale and alter the contextual appearance of the data. [EXAMPLE HERE] One common graphical display choice is the use of log scales used to display data over several orders of magnitude within one graph. Logarithms make multiplicative relationships additive, providing an elegant way to span many orders of magnitude, to show elasticities and other proportional

CONTACT Emily A. Robinson. Email: emily.robinson@huskers.unl.edu, Reka Howard. Email: rekahoward@unl.edu, Susan VanderPlas. Email: susan.vanderplas@unl.edu

changes, and to linearize power laws (Menge et al. 2018). When presenting log-scaled data, it is possible to use either untransformed values (for example, values of 1, 10 and 100 are equally spaced along the axis) or log-transformed values (for example, 0, 1, and 2). This is the difference between a transformed scale transformed values; it's not the difference between untransformed values and transformed values. see https://r4ds.had.co.nz/graphics-for-communication.html#scales for an explanation of scales, but I'm still looking for a perfect reference for transformations at the value vs. scale level. We have recently experienced the benefits and pitfalls of using log-scales as covid-19 dashboards displayed case count data on both the log and linear scale (Fagen-Ulmschneider 2020). INSERT BENEFITS AND PITFALLS OF LOG SCALES HERE. While COVID-19 is the most well known example, log-scales have been used to display data in ecological research, etc. PUT OTHER AREAS HERE.

Previous research suggests our perception and mapping of numbers to a number line proceeds logarithmically at first and transitions to linear later in development, after formal schooling begins. The logarithmic mapping is more noticeable in larger numbers as the transition to linear mapping occurs first for small numbers in young children and later for larger numbers (Varshney and Sun 2013; Siegler and Braithwaite 2017; Dehaene et al. 2008).Invert this sentence, because right now it's a bit confusing... e.g. In young children, the transition to linear scales occurs first in small numbers (e.g. 1-10) and then gradually expands to higher orders of magnitude; thus, the logarithmic intuition about numbers in children is often more noticeable on scales in the thousands to hundreds of thousands. Early studies explored the estimation and prediction of exponential growth. Findings indicate that growth is underestimated when presented both numerically and graphically but that numerical estimation is more accurate than graphical estimation for exponential curves. While prior contextual knowledge or experience with exponential growth does not improve estimation, previous instruction on exponential growth reduces the underestimation by adjusting the initial starting value but not adjusting their perception of growth parameter (Wagenaar and Sagaria 1975; Jones 1977). Estimation was shown to improve when subjects were presented with decreasing exponential functions (Timmers and Wagenaar 1977). Jones (1977), Wagenaar and Timmers (1978), and Jones (1979) propose competing polynomial models for the perception and extrapolation of exponential series. We hypothesize that estimation is a two-stage process. First, we identify the type of curve and direction and then use that information for prediction (Best, Smith, and Stubbs 2007).

This paper aims to investigate the use of log scales to display exponentially increasing data. We hypothesize log scales should make it much easier to estimate the growth paramter since we estimate slopes relatively accurately, resolving much of the difficulty with exponential estimation (Mosteller et al. 1981). In Menge et al. (2018), ecologists were surveyed to determine how often ecologists encounter log-scaled data and how well ecologists understand log-scaled data. Participants were presented two relationships displayed on linear-linear scales, log-log scales with untransformed values, or log–log scales with log-transformed values. Menge et al. (2018) propose three types of misconceptions participants encountered when presented data on log-log scales: 'hand-hold fallacy', 'Zeno's zero fallacy', and 'watch out for curves fallacies'.

Might be easier to start with graphs are statistics, but that to evaluate a graph we have to run our statistic through a visual evaluation - a person. Then you can explain that with the same data, if two different methods of presenting the data result in qualitatively different results re. evaluation by a human, then it is possible to develop recommendations based on the visual inference experiment. Recent graphical experiments have utilized statistical lineups to quantify the perception of graphical

design choices(VanderPlas and Hofmann 2017). Statistical lineups provide an elegant way of combining perception and statistical signficiance by valdiating the findings from a graphical experiment (Buja et al. 2009; Wickham et al. 2010; Hofmann et al. 2012; Majumder, Hofmann, and Cook 2013; VanderPlas and Hofmann 2017). 'Lineups' are named after the 'police lineup' of criminal investigations where witnesses are asked to identify the criminal from a set of individuals. Similarly, a statistical lineup is a plot consisting of smaller panels of plots in which the viewer is asked to identify the plot of the real data from among a set of decoys, the null plots. A statistical lineup typically consists of 20 panels - 1 target panel and 19 null panels (INSERT EXAMPLE). If the viewer can identify the target panel embeded within the set of null panels, this suggests that the real data is interesting or has unqiue properties. Crowd sourcing websites such as Amazon Mechanical Turk, Reddit, and Proflic allow us to collect responses from multiple viewers. VanderPlas et al. (n.d.) provides an approach for calculating visual p-values utilizing a 'rorschach' lineup which consists soley of null panels.

In this paper, we use statistical lineups to test human subjects ability to differentiate between increasing exponential data with differing growth rates displayed on both the linear scale and log scale.

2. Data Generation

The most common type of lineup used in graphical experiments is a standard lineup containing one "target" dataset embeded within a set of null datasets. One way to generate the null datasets when working with real data is through the use of permutation. In this study, both the target and null datasets were generated by simulating data from an exponential model with differing parameters. This experiment was designed to test a participants ability to differentiate between different rates of exponential growth on both the log and linear scales. In order to gaurantee the simulated data spans the same range of values, we implemented a range constraint of $y \in [10, 100]$ and a domain constraint of $x \in [0, 20]$ with $N = 50$ points randomly assigned throughout the domain and mapped to the y-axis using the exponential model with the selected paramters. These constraints provide some assurance that participants who select the target plot are doing so because of their visual perception differentiating between curvature or slope rather than different starting or ending values.

2.1. *Exponential Model*

We simulated data based on a 3-parameter exponential model with multiplicative errors. This model has a β parameter to reflect the rate of growth and amount of curvature and σ^2 to reflect the amount of variability around the exponential growth curve. The parameters α and θ are adjusted based on β and σ^2 to gauranee the range and domain constraints are met. The model generated $N = 50$ points $(x_i, y_i), i = 1, ..., N$ where x and y have an increasing exponential relationship. The data was simulated heuristically by the following procedures:

Algorithm 2.1.1: Paremeter Estimation
Input Parameters: domain $x \in [0, 20]$, range $y \in [10, 100]$, midpoint x_{mid}.
Output: estimated model parameters $\hat{\alpha}, \hat{\beta}, \hat{\theta}$

1. Determine the $y = -x$ line scaled to fit the assigned domain and range.

2. Map the values $x_{mid} - 0.1$ and $x_{mid} + 0.1$ to the $y = -x$ line for two additional points.
3. From the set points (x_k, y_k) for $k = 1, 2, 3, 4$, obtain the coefficients from the linear model $\ln(y_k) = b_0 + b_1 x_k$ to obtain starting values - $\alpha_0 = e^{b_0}, \beta_0 = b_1, \theta_0 = 0.5 \cdot \min(y)$
4. Using the 'nls()' function from the 'stats' package in Rstudio and the starting parameter values - $\alpha_0, \beta_0, \theta_0$ - fit the nonlinear model, $y_k = \alpha \cdot e^{\beta \cdot x_k} + \theta$ to obtain estimated parameter values - $\hat{\alpha}, \hat{\beta}, \hat{\theta}$.

Alogrithm 2.1.2: Exponential Simulation

Input Paremeters: sample size $N = 50$, estimated parameters $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\theta}$, σ standard deviation from the exponential curve.

Output Parameters: N points, in the form of vectors $\mathbf{x}$ and $\mathbf{y}$

1. Generate $\tilde{x}_j, j = 1, ..., N \cdot \frac{3}{4}$ as a sequence of evenly spaced points in $[0, 20]$. This ensures the full domain of x is used, fulfilling the constraints of spanning the same domain and range for each parameter combination.
2. Obtain $\tilde{x}_i, i = 1, ...N$ by sampling $N = 50$ values from the set of $\tilde{x}_j$ values. This gaurantees some variability and potential clustring in the exponential growth curve disrupting the perception due to continuity of points.
3. Obtain the final x_i values by jittering $\tilde{x}_i$.
4. Calculate $\tilde{\alpha} = \frac{\hat{\alpha}}{e^{\sigma^2/2}}$. This ensures that the range of simulated values for different standard devaition parameters has an equal expected value for a given rate of change due to the non-constant variance across the domain.
5. Generate $y_i = \tilde{\alpha} \cdot e^{\hat{\beta} x_i + e_i} + \hat{\theta}$ where $e_i \sim N(0, \sigma^2)$.

2.2. *Parameter Selection*

The exponential model provides the base for this graphical experiment. We manipulate the midpoint, x_{mid}, and in turn the estimated parameters to control the amount of curvature present in the data and the error standard deviation, σ, to control the amount of deviation from the exponential curve. We selected three midpoints corresponding to difficulty levels easy (obvious curvature), medium (noticable curvature), and hard (almost linear) along with a sensible choice of standard deviation, σ. (INSERT TABLE WITH PARAMETER SELECTIONS HERE). The midpoints and standard deviation combinations were chosen similar to in VanderPlas and Hofmann (2017). For each level of difficulty, we simulated 1000 datasets of (x_{ij}, y_{ij}) points for $i = 1, ..., 50$ and $j = 1...10$. Each generated x_i point from *Algorithm 2.1.2* was replicated 10 times. Then the lack of fit statistic (LOF) was computed for each simulated dataset by calculating the deviation of the data from a linear line. Plotting the density curves of the LOF statistics for each level of difficulty choice allows us to evalute the ability of differentiating between the difficulty levels and thus detecting the target plot. In figure [INSERT DENSITY CURVE HERE], we can see the densities of each of the three difficulty levels. While the LOF statistic provides us a numerical value for discriminating between the difficulty levels, we cannot directly realte this to the perceptual discriminability.

2.3. *Lineup Setup*

There were a total of three parameter combinations corresponding to the three difficulty levels - easy (obvious curvature), medium (noticable curvature), and hard (almost linear). The lineup plots were generated by mapping simulating data corresponding to difficulty level A to a scatterplot while the null plots were generated by mapping simulated data corresponding to difficulty level B to a scatterplot. For exmaple, a target plot with simulated data following an increasing exponential curve with obvious curvature is embeded within null plots with simulated data following an increasing exponential that is almost linear (i.e. Easy-Hard). By our constraints, the target plot and null plots will span a similar domain and range. There are a total of 6 (i.e. 3 choose 2) lineup parameter combinations. Two sets of each lineup parameter combination were simulated (total of 12 test datasets) and plotted on both the linear and the log scale (total of 24 test lineup plots). It is worth noting that there were also three rorschach lineup parameter combinations (e.g. Easy-Easy) and that each of these also had two sets of datasets simulated and plotted on both scales (12 rorscahch lineup plots).

2.4. *Study Design*

Each participant was shown a total of thirteen lineup plots (twelve) test lineup plots and one rorschach lineup plot). Participants were randomly assigned one of the two replicate datasets for each of the six unique lineup parameter combinations. For each assigned test dataset, the participant was shown the lineup plot corresponding to both the linear scale and the log scale. For the additional rorschach lineup plot, participants were randomly assigned a rorschach dataset on either the linear or the log scale. The order of the thirteen lineup plots shown was randomized for each participant.

2.5. *Participant Recruitment*

Participants above age 19 were recruited from Reddit's R Visualization community. Since participants recruitted on Reddit were not compensated for their time, most participants have an interest in data visualization research. Previous literature suggests that prior mathematical knowledge or experience with exponential data is not associated with the outcome of graphical experiments (SITE THIS!). Participants were then directed to complete the experimental task available at LINK TO SHINY APP HERE.

2.6. *Task Description*

Participants were shown a series of twelve test lineup plots and asked to identify the plot that was most different from the others. On each plot, participants were asked to justify their choice and provide their level of confidence in their choice. The goal of this experimental task is to test an individuals ability to perceptually differentiate exponentially increasing data with differing rates of change on both the linear and log scale. In Best, Smith, and Stubbs (2007), the authors explored whether descrimination between curve types is possible. They found that accuracy higher when nonlinear trends presented (e.g. it's hard to say something is linear, but easy to say that it isn't) and that accuracy higher with low additive variability.

3. Results

Participant data was anlalyzed using the Glimmix Procedure in SAS 9.4. Each lineup plot evaluated was assigned a value based on the participant response (correct = 1, not correct = 0). The binary response was analyzed using generalized linear mixed model following a binomial distribution.

3.1. *Curvature Differentiation*

3.2. *Linear vs Log*

3.3. *Participant Reasoning*

4. Discussion and Conclusion

5. Future Research

In this study, we discovered that differentiation between data following exponentially increasing trends with differing growth rates is FINISH THIS.

Further experimentation is necessary to test an individual's ability to make predictions for exponentially increasing data. Previous literature suggests that we tend to underestimate predictions of exponentially increasing data.(Jones 1979, 1977; Wagenaar and Timmers 1978). (Mosteller et al. 1981) designed and carried out an empirical investigation to explore properties of lines fitted by eye. The researchers found that students tended to fit the slope of the first principal component or major axis (the line that minimizes the sum of squares of perpendicular rather than vertical distances) and that students who gave steep slopes for one data set also tended to give steep slopes on the others. Interestingly, the individual-to-individual variability in slope and in intercept was near the standard error provided by least squares. A similar graphical taxk is used in the New York Times "You Draw It" page asking readers to test their knowledge by using their curser to estimate values of a certain topic under different political administrations or over different years (CITE THIS).

In addition to differentiation and prediction of exponentially increasing data, it is of interest to test an individuals ability to translate a graph of exponentially increasing data into real value quantities and extend their estimations by making comparisons. (Friel, Curcio, and Bright 2001) emphasize the importance of graph comprehension proposing that the graph construction plays a role in the ability to read and interpret graphs.

Supplementary Materials

References

Best, Lisa A., Laurence D. Smith, and D. Alan Stubbs. 2007. "Perception of Linear and Nonlinear Trends: Using Slope and Curvature Information to Make Trend Discriminations." *Perceptual and Motor Skills* 104 (3): 707–721. Publisher: SAGE Publications Inc, Accessed 2020-07-06. https://doi.org/10.2466/pms.104.3.707-721.

Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wickham. 2009. "Statistical inference for exploratory data analysis and model diagnostics." *Philosophical Transactions of the Royal Society A: Mathematical,*

Physical and Engineering Sciences 367 (1906): 4361–4383. Accessed 2020-10-06. https://royalsocietypublishing.org/doi/10.1098/rsta.2009.0120.

Charlotte, Lisa. 2020. "You've informed the public with visualizations about the coronavirus. Thank you." Jul. https://blog.datawrapper.de/datawrapper-effect-corona/.

Dehaene, Stanislas, Véronique Izard, Elizabeth Spelke, and Pierre Pica. 2008. "Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigene Cultures." *Science* 320 (5880): 1217–1220. 00651 Publisher: American Association for the Advancement of Science Section: Report, Accessed 2020-05-19. https://science.sciencemag.org/content/320/5880/1217.

Fagen-Ulmschneider, Wade. 2020. "91-DIVOC." http://91-divoc.com/pages/covid-visualization/.

Friel, Susan N., Frances R. Curcio, and George W. Bright. 2001. "Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications." *Journal for Research in Mathematics Education* 32 (2): 124. Accessed 2020-05-29. https://www.jstor.org/stable/749671?origin=crossref.

Hofmann, Heike, Lendie Follett, Mahbubul Majumder, and Dianne Cook. 2012. "Graphical Tests for Power Comparison of Competing Designs." *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2441–2448. Accessed 2020-04-06. http://ieeexplore.ieee.org/document/6327249/.

Jones, Gregory V. 1977. "Polynomial perception of exponential growth." *Perception & Psychophysics* 21 (2): 197–198. Accessed 2020-06-25. http://link.springer.com/10.3758/BF03198726.

Jones, Gregory V. 1979. "A generalized polynomial model for perception of exponential series." *Perception & Psychophysics* 25 (3): 232–234. Accessed 2020-05-19. http://link.springer.com/10.3758/BF03202992.

Majumder, Mahbubul, Heike Hofmann, and Dianne Cook. 2013. "Validation of Visual Statistical Inference, Applied to Linear Models." *Journal of the American Statistical Association* 108 (503): 942–956. Accessed 2020-02-28. http://www.tandfonline.com/doi/abs/10.1080/01621459.2013.808157.

Menge, Duncan N. L., Anna C. MacPherson, Thomas A. Bytnerowicz, Andrew W. Quebbeman, Naomi B. Schwartz, Benton N. Taylor, and Amelia A. Wolf. 2018. "Logarithmic scales in ecological data presentation may cause misinterpretation." *Nature Ecology & Evolution* 2 (9): 1393–1402. Accessed 2020-08-18. http://www.nature.com/articles/s41559-018-0610-7.

Mosteller, Frederick, Andrew F. Siegel, Edward Trapido, and Cleo Youtz. 1981. "Eye Fitting Straight Lines." *The American Statistician* 35 (3): 150–152. Accessed 2020-05-29. http://www.tandfonline.com/doi/abs/10.1080/00031305.1981.10479335.

Siegler, Robert S., and David W. Braithwaite. 2017. "Numerical Development." *Annual Review of Psychology* 68 (1): 187–213. Accessed 2020-05-19. http://www.annualreviews.org/doi/10.1146/annurev-psych-010416-044101.

Timmers, Han, and Willem A. Wagenaar. 1977. "Inverse statistics and misperception of exponential growth." *Perception & Psychophysics* 21 (6): 558–562. Accessed 2020-07-06. http://link.springer.com/10.3758/BF03198737.

Unwin, Anthony. 2020. "Why is Data Visualization Important? What is Important in Data Visualization?" *Harvard Data Science Review* Accessed 2020-04-27. https://hdsr.mitpress.mit.edu/pub/zok97i7p.

VanderPlas, Susan, and Heike Hofmann. 2017. "Clusters Beat Trend!? Testing Feature Hierarchy in Statistical Graphics." *Journal of Computational and Graphical Statistics* 26 (2): 231–242. Accessed 2020-02-28. https://www.tandfonline.com/doi/full/10.1080/10618600.2016.1209116.

VanderPlas, Susan, Christian Rottger, Dianne Cook, and Heike Hofmann. n.d. "Statistical Significance Calculations for Scenarios in Visual Inference." Preprint, Accessed 2020-09-29. https://github.com/srvanderplas/visual-inference-alpha.

Varshney, Lav R., and John Z. Sun. 2013. "Why do we perceive logarithmically?" *Significance*

10 (1): 28–31. Accessed 2020-05-07. http://doi.wiley.com/10.1111/j.1740-9713.2013.00636.x.

Wagenaar, W. A., and H. Timmers. 1978. "Extrapolation of exponential time series is not enhanced by having more data points." *Perception & Psychophysics* 24 (2): 182–184. Accessed 2020-05-19. http://link.springer.com/10.3758/BF03199548.

Wagenaar, William A., and Sabato D. Sagaria. 1975. "Misperception of exponential growth." *Perception & Psychophysics* 18 (6): 416–422. Accessed 2020-07-07. http://link.springer.com/10.3758/BF03204114.

Wickham, Hadley, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. "Graphical inference for infovis." *IEEE Transactions on Visualization and Computer Graphics* 16 (6): 973–979.