Human Perception of Statistical Charts: An Introduction to Graphical Testing Methods

UNL Nerd Nite @ Saro Cider

October 13, 2021

Emily A. Robinson

Department of Statistics, University of Nebraska - Lincoln

emily.robinson@huskers.unl.edu

www.emilyarobinson.com

earobinson95

Slides: https://bit.ly/3msOdKX

Thank you, everyone for coming! I am a PhD candidate in the Department of Statistics at the University of Nebraska - Lincoln. I will be presenting on human perception of statistical charts and giving an overview of current graphical testing methods then introduce the current research I am conducting in graphical testing.

Introduction to Graphics

Data visualization is defined as the art of drawing **graphical charts** in order to display data (Unwin, 2020).

What are graphics useful for? (Lewandowsky and Spence, 1989)

Data cleaning.

Exploring data structure.

Communicating information.

Who uses graphics?

- Governments (Harms, 1991; Playfair, 1801; Walker, 2013).
- Companies (Chandar, Collier, and Miranti, 2012; Yates, 1985).
- News sources and mass media (Aisch, Cohn, Cox, et al., 2016).
- Scientific publications (Gouretski and Koltermann, 2007).

To get started, we are first going to lay the foundation of graphics. Data visualization has become central tool in modern data science and statistics. Unwin 2020 defines data visualization as the art of drawing graphical charts in order to display data.

Graphics are useful for data cleaning, exploring data structure, and communicating information.

In the 18th and 19th century, governments began using graphics to understand population and economic interests. In the 20th century, we saw companies using graphics to understand the inner workings of their business and support their business decision. We often see news source displaying graphics of weather forecasts such as hurricane trajectories. Today, we see graphics everywhere from scientific journals to mass media in the newspapers, TV, and internet.

Despite the popularity of graphics, we are too accepting of them as default without asking critical questions about the graphics we create or view (Unwin, 2020). We must begin asking ourselves **How effective is this graph at communicating useful information?**

Higher quality of technology has influenced the creation, replication, and complexity of graphics. We now have an infinitely many number design choices:
+ variables displayed, type of graphic, size of graphic, aspect ratio, colors, symbols, scales, limits, ordering of categorical variables

There is a need for an established set of concepts and terminology to build their graphics from so they can actively choose which of many possible graphics to draw in order to ensure their charts are effective at communicating the intended result.

Grammar of Graphics (Wilkinson, 2013)

Graphics are viewed as a mapping **from variables** in the data set **to visual attributes** on the chart.

Building a masterpiece, by Allison Horst

Graphics are viewed as a mapping **from variables** in the data set **to visual attributes** on the chart.

+ **aesthetics:** links between data variables and graphical features (position, color, shape, size)
+ **layers:** geometric elements (points, lines, rectangles, text, …)
+ **transformations:** transformations specify a functional link between the data and the displayed information (identity, count, bins, density, regression). Transformations act on the variables.
+ **scales:** scales map values in data space to values in the aesthetic space. Scales change the coordinate space of an aesthetic, but don't change the underlying value (so the change is at the visual level, not the mathematical level).
+ **coordinate system:** e.g.polar or Cartesian
+ **faceting:** facets allow you to split plots by other variables to produce many sub-plots.
+ **theme:** formatting items, such as background color, fonts, margins…

Software, such as Hadley Wickham's ggplot2, aims to implement the framework of creating charts and graphics as the grammar of graphics recommends.

Testing Statistical Graphics

Evaluate design choices and understand cognitive biases through the use of **visual tests**.

Could ask participants to:

📊 identify differences in graphs.

🕮 read information off of a chart accurately.

🌎 use data to make correct real-world decisions.

✏ predict the next few observations.

One way we can evaluate these design choices through the use of graphical tests.

Could ask participants to:

- identify differences in graphs.
- read information off of a chart accurately.
- use data to make correct real-world decisions.
- predict the next few observations.

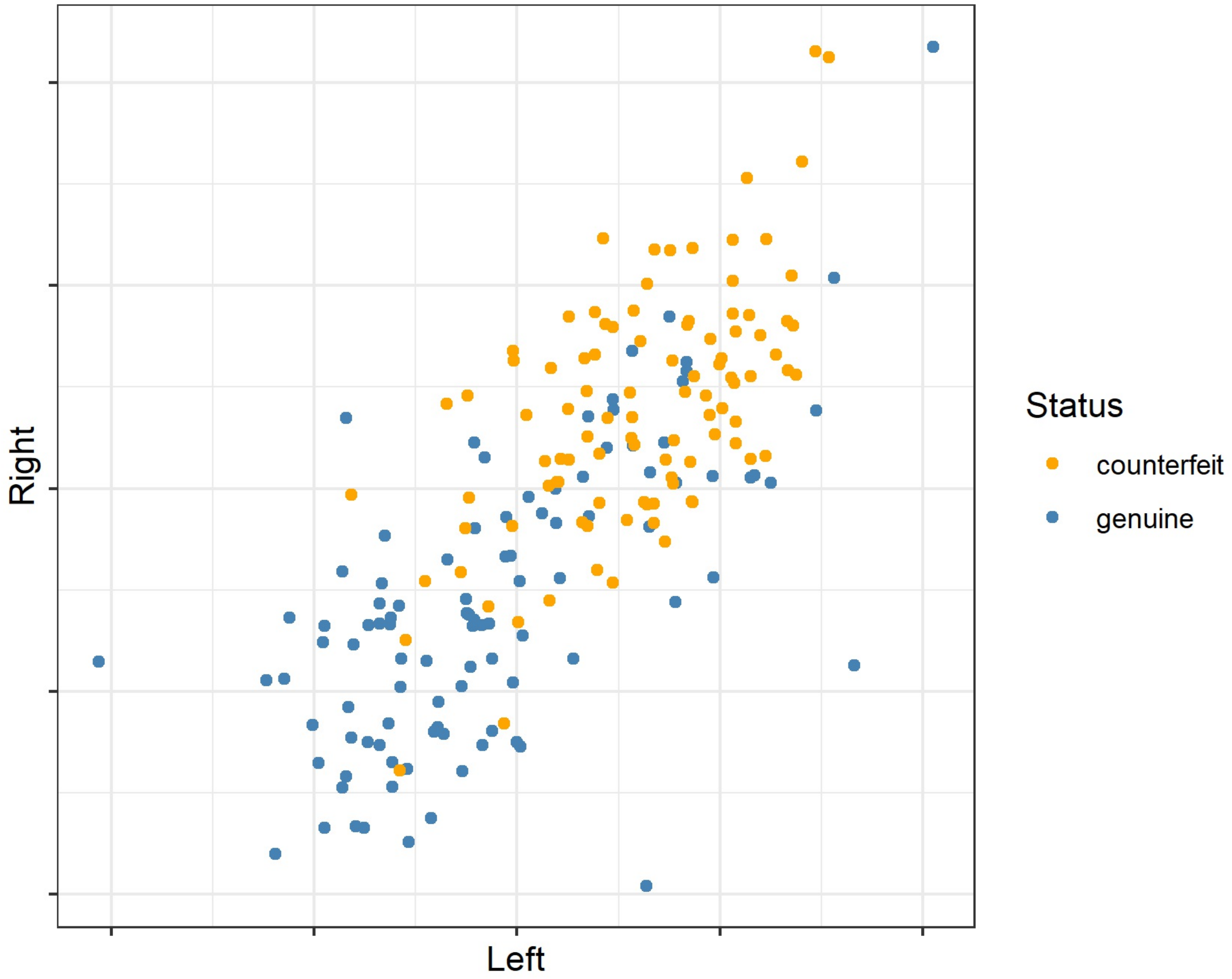
All of these types of tests require different levels of use and manipulation of the information presented in the chart.

Lineup Protocal

(Buja et. al, 2009)

Efforts in the field of graphics have developed graphical testing tools and methods such as the lineup protocol to provide a framework for inferential testing.

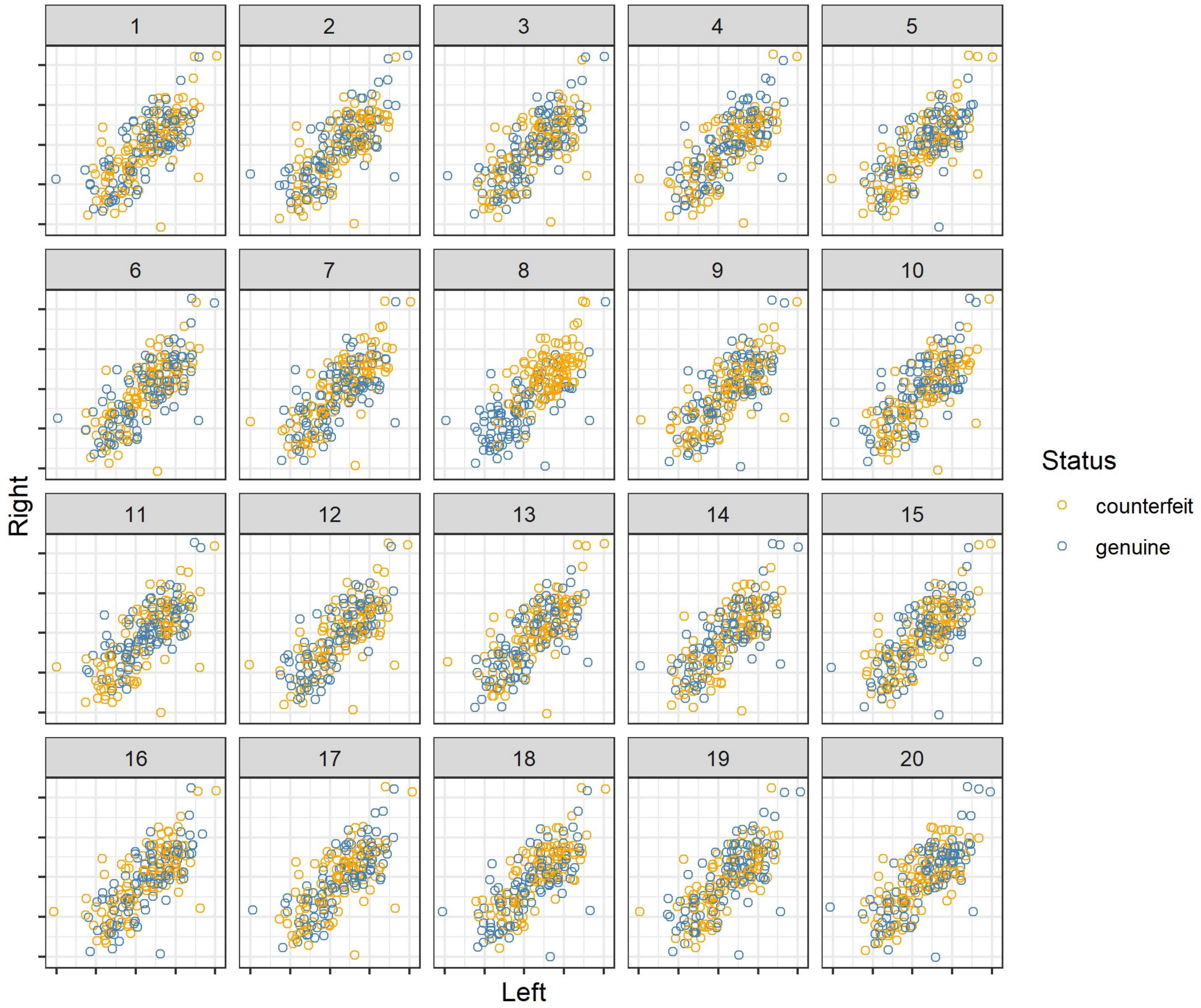
When doing exploratory data analysis, how do we know if what we see is actually there?



When inspecting a plot, how do we know if what we are seeing is actually there?

Lineup Protocal (Buja et. al, 2009)

Embed a **target plot** (actual data) in a lineup of **null plots** (randomly permuted data sets).

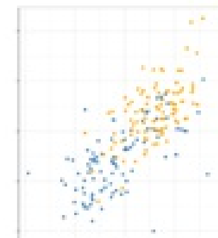
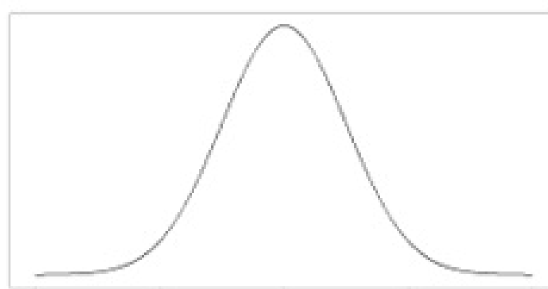
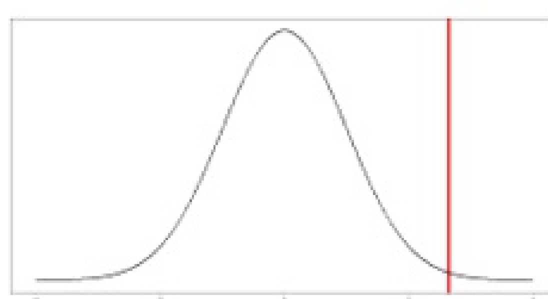
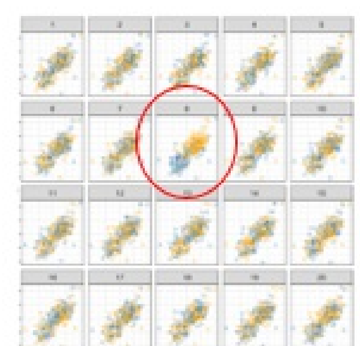


One way of answering this question is to embed the true data plot (called target plot) into a set of randomly permuted data sets (called null plots). This is what we call a lineup.

This is similar to the law-enforcement procedure to line up a suspect among a set of innocents to check if a victim can identify the suspect as the perpetrator of the crime.

Here, visual evaluation of the lineup is conducted by a person. If the viewers detect the target plot, we can conclude the plots are distinguishable.

The lineup protocol is one such example of the development of tools designed for statistical graphical testing. The advancement of graphing software provides the tools necessary to develop new methods of testing graphics.

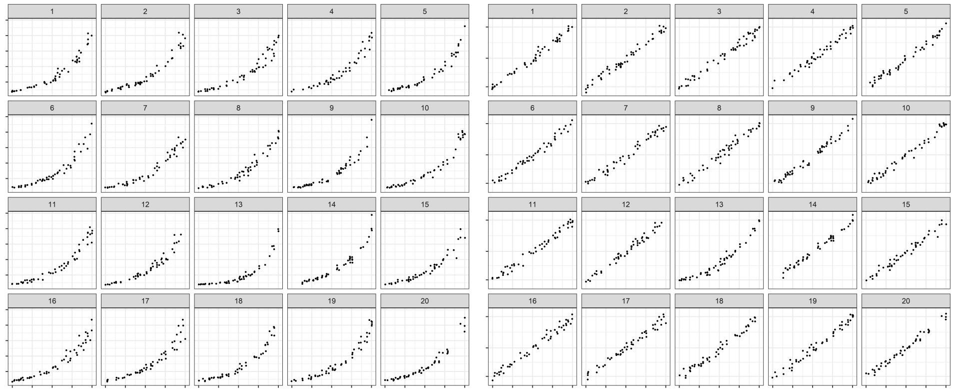
For real data set $\boldsymbol{y}$	
Traditional Inference	**Visual Inference**
Calculate observed test statistic. $Z_O = \frac{\bar{x} - \mu}{\sigma}$	Plot the real data (visual statistic). 
Determine null distribution. $Z \sim N(0,1)$ 	Embed in a lineup of null plots. 
Reject if observed statistic is beyond a cutoff. $Z_O > 2.64$ 	Evaluated by human viewers. Reject if target panel is correctly identified. 

The main idea behind visual inference is that graphs are visual statistics or summaries of the data sets generated by mathematical functions. In a standard statistical analysis, a test statistic is generated from the data set and compared to the null distribution of that test statistic. Similarly, the visual statistic (target plot) is compared by a human viewer to other plots generated under the assumption of the null.

The main benefit of using visual inference is that visual tests tend to be more comprehensive. Since individuals are being asked to select one or more plots from a the lineup which are different but the difference is left unspecified the individual eye is actually spotting many differences at once. An equivalent numerical assessment may involved multiple tests using different test statistics.

- Statistical inference for exploratory data analysis and model diagnostics (Buja, Cook, Hofmann, et al., 2009)
- Validation of Visual Statistical Inference, Applied to Linear Models (Majumder, Hofmann, and Cook, 2013)
- Human Factors Influencing Visual Statistical Inference (Majumder, Hofmann, and Cook, 2014)
- Variations of *Q–Q* Plots: The Power of Our Eyes! (Loy, Follett, and Hofmann, 2016)
- Spatial Reasoning and Data Displays (VanderPlas and Hofmann, 2015)
- Clusters beat trend!? testing feature hierarchy in statistical graphics (VanderPlas and Hofmann, 2017)
- Statistical Significance Calculations for Scenarios in Visual Inference (VanderPlas, Röttger, Cook, et al., 2021)

- **Dissertation work:** Perception of log scales



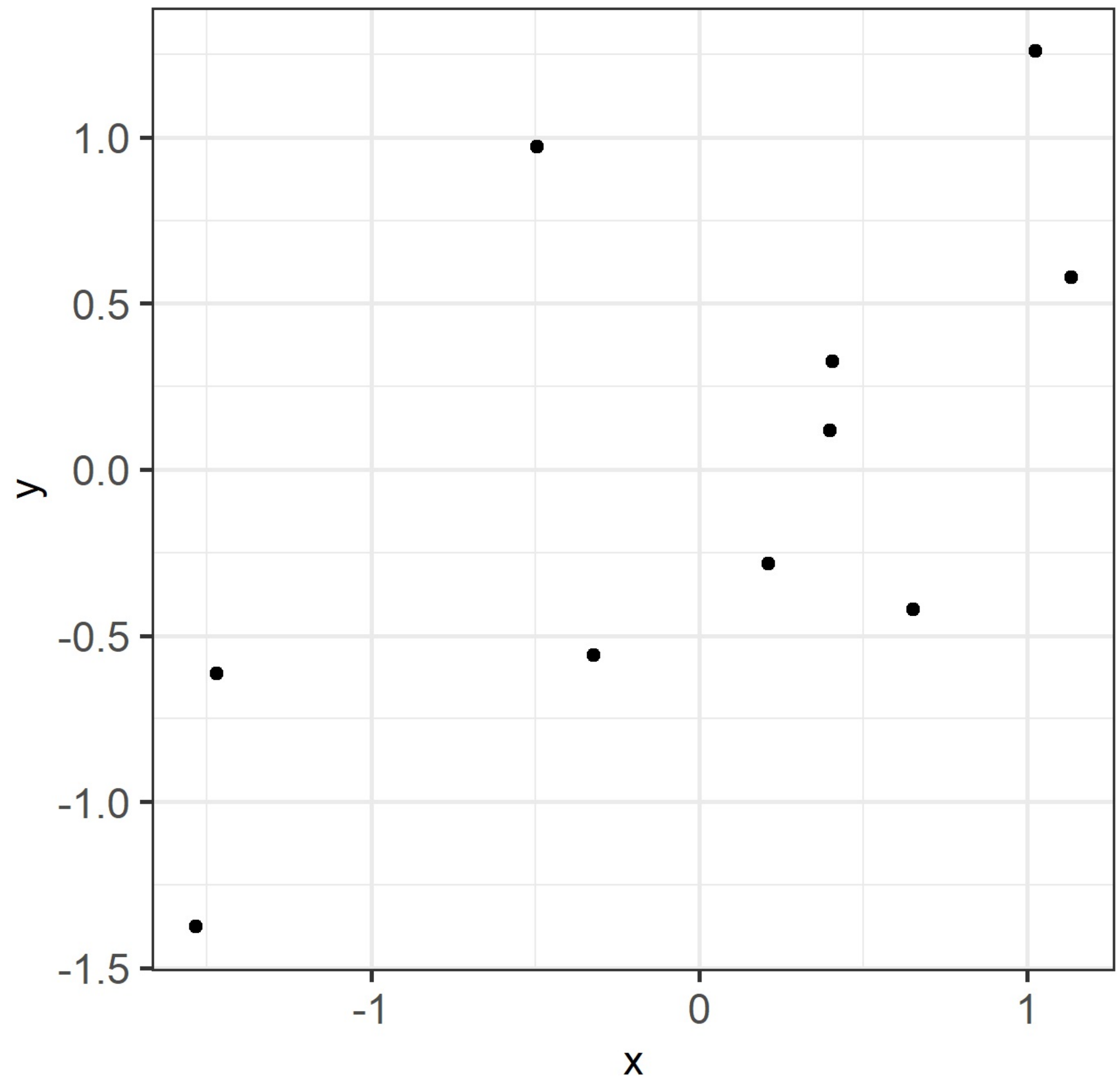
Clusters beat trend!? testing feature hierarchy in statistical graphics `r Citep(bib[[c("vanderplas2017clusters")]])`
+ Introduces perceptual principles such as preattentive features and gestalt heuristics
+ Discusses the design and results of a factorial experiment examining the effect of plot aesthetics such as color and trend lines on participants' assessment of ambiguous data displays.
+ Strongly suggests that plot aesthetics have a significant impact on the perception of important features in data displays.

As part of my dissertation work, I have conducted a study examining human perception of logarithmic scales. Here we see an example of the first part of the study utilizing lineups to test our ability to perceptually differentiate between two exponentially increasing trends shown on both the linear and log scale. Notice how it is much easier to pick out panel 13 as being most different when displayed on the log scale than on the linear scale.

'You Draw It'

Linear Regression

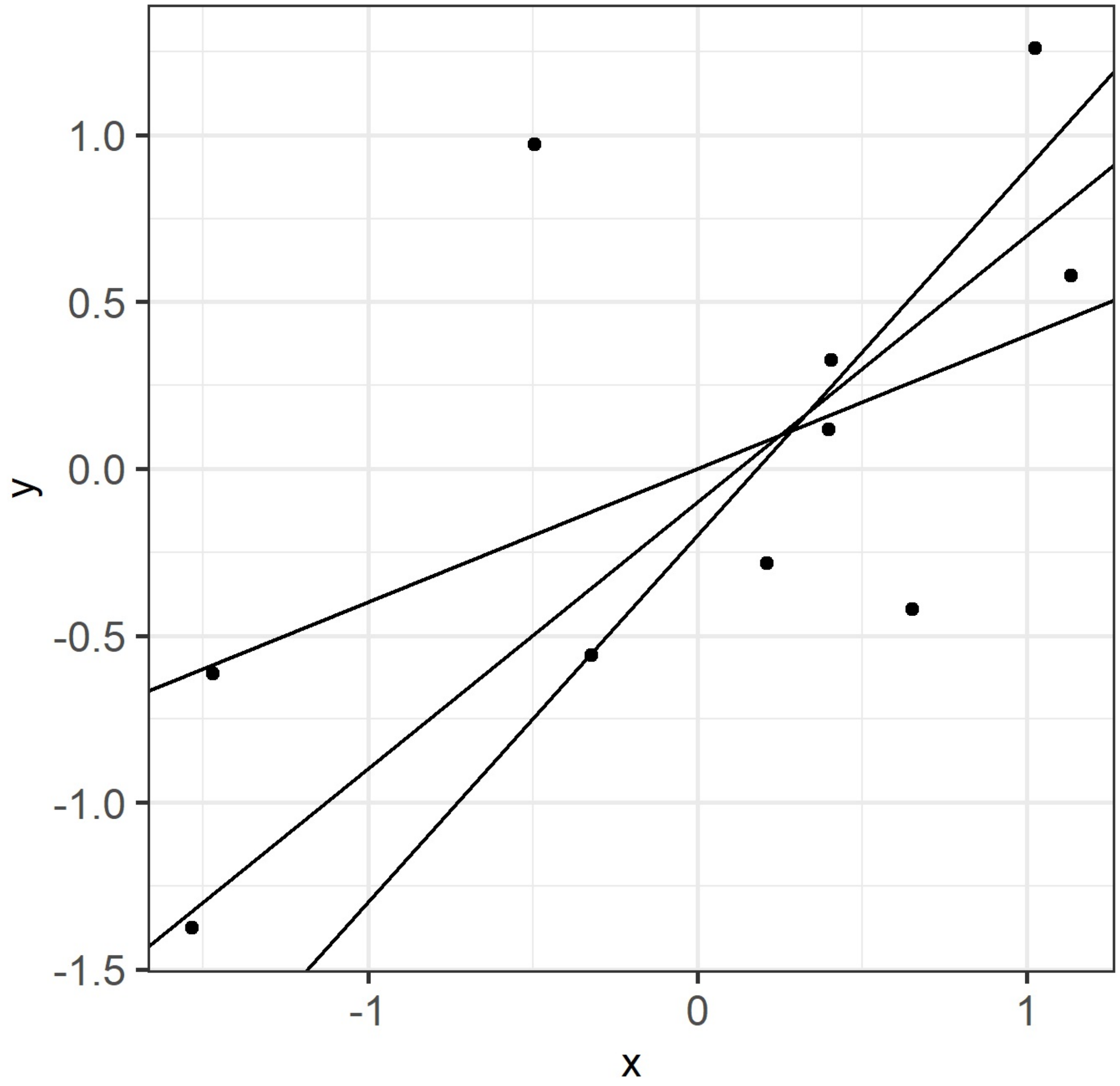
The principle of simple linear regression is to find the line (i.e., determine its equation) which passes as close as possible to the observations, that is, the set of points.



Linear regression is a statistical approach that allows to assess the linear relationship between two quantitative variables.

Linear Regression

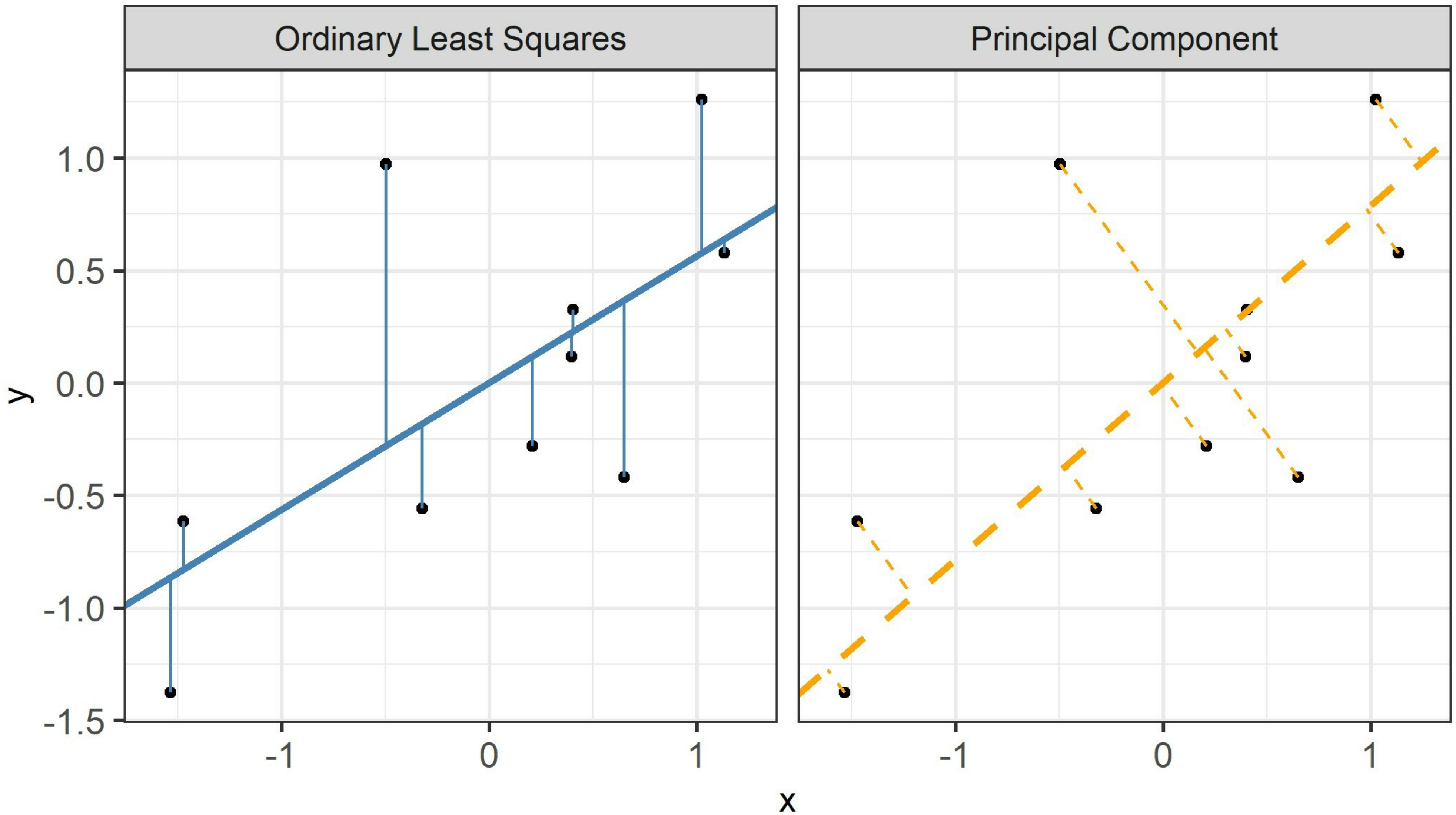
The principle of simple linear regression is to **find the line** (i.e., determine its equation) which passes as close as possible to the observations, that is, the set of points.



The principle of simple linear regression is to **find the line** (i.e., determine its equation) which passes as close as possible to the observations, that is, the set of points.

Linear Regression

The principle of simple linear regression is to **find the line** (i.e., determine its equation) **which passes as close as possible to the observations**, that is, the set of points.



Big Idea: How do statistical regression results compare to intuitive, visually fitted results?

We are going to focus on two regression lines determined by ordinary least squares regression and regression based on the principal axis. The figure illustrates the difference between an OLS regression line which minimizes the vertical distance of points from the line and a regression line based on the principal axis (Principal Component) which minimizes the Euclidean distance of points (orthogonal) from the line. This is what we refer to as “ensemble perception” indicating the visual system can compute averages of various features in parallel across the items in a set (in this case, over the x and y-axes).

Big Idea: How do statistical regression results compare to intuitive, visually fitted results?

Mosteller, Siegel, Trapido, et al. (1981)

- **Big Idea:** Students fitted lines by eye to four sets of points.
- **Method:** 8.5 x 11 inch transparency with a straight line etched across the middle.
- **Sample:** 153 graduate students and post docs in Introductory Biostatistics.
- **Experimental Design:** Latin square.
- **Findings:** Students tended to fit the slope of the first principal component.

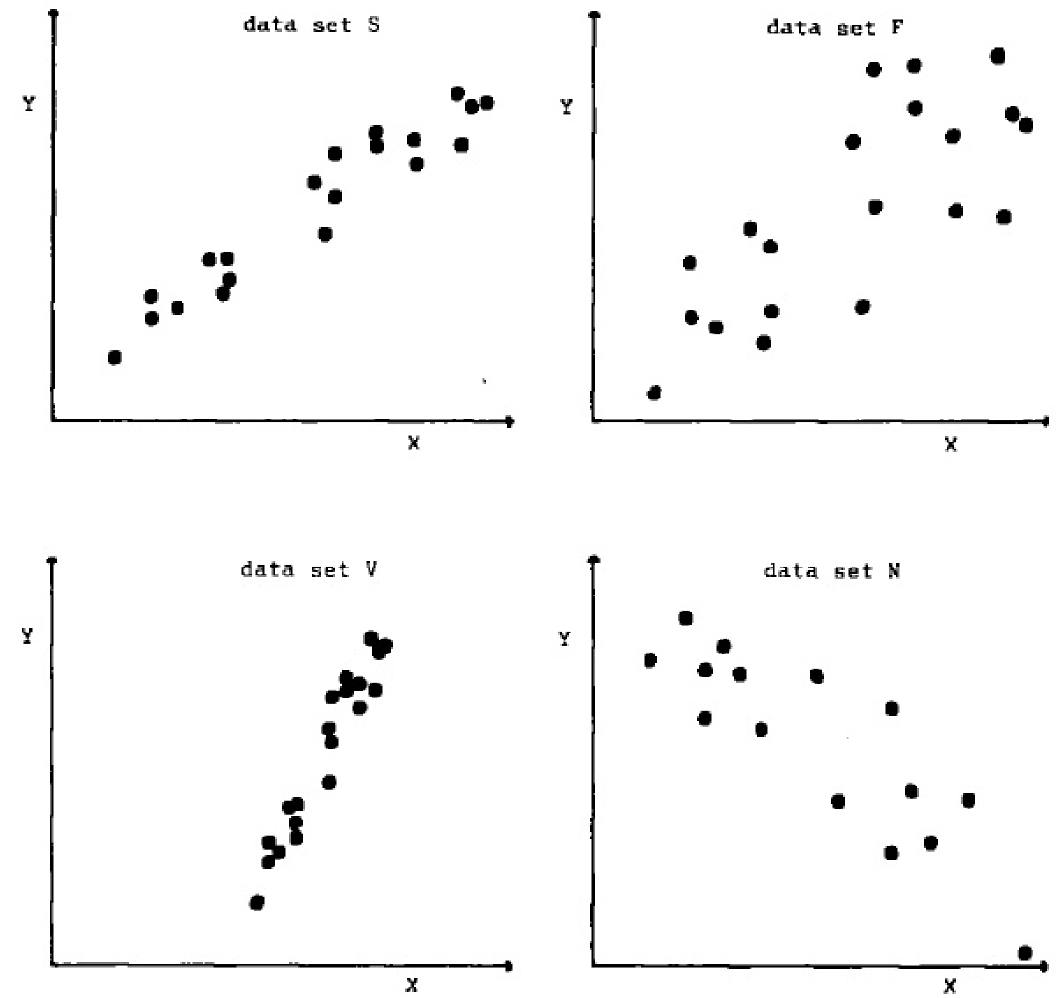


Figure 1. The Data Sets of S, F, V, and N

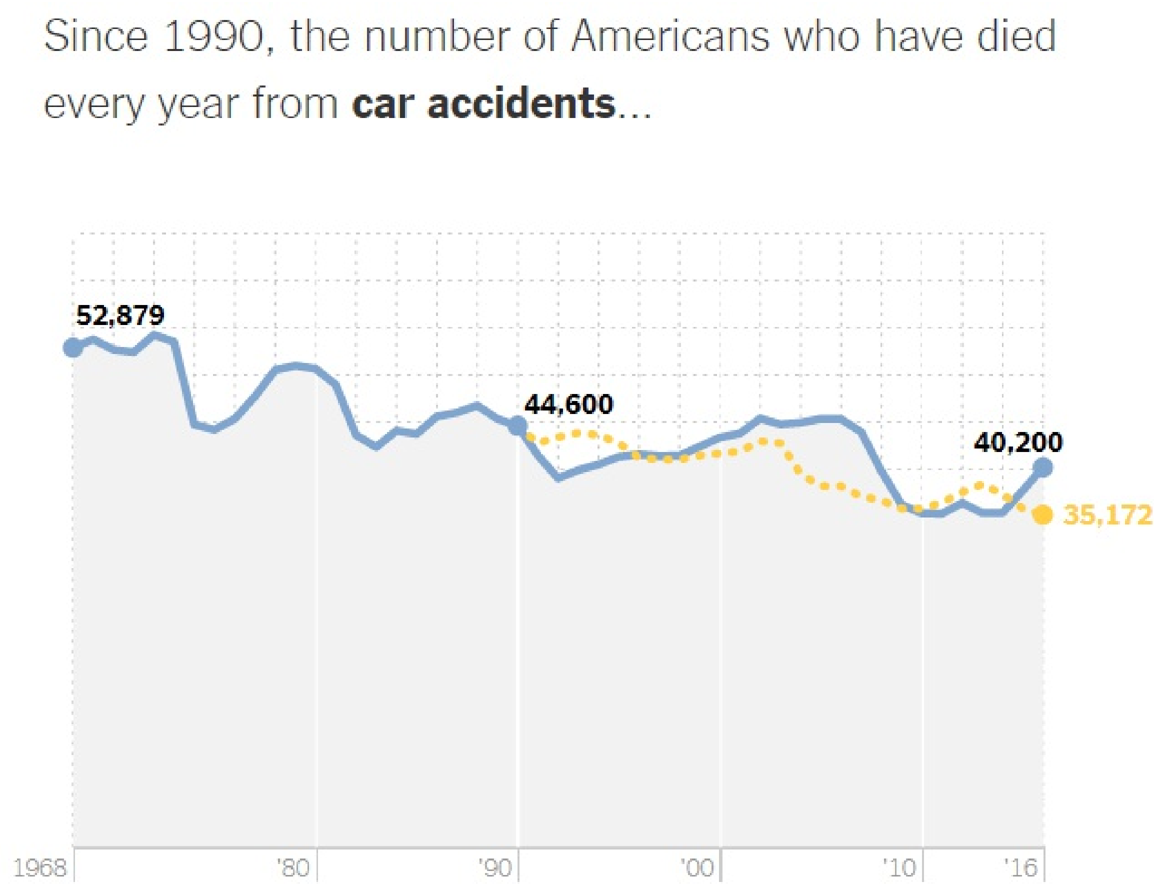
I want to introduce a study conducted in 1981 called Eye Fitting Straight Lines by Mosteller et al. In this study:

+ Students fitted lines by eye to four sets of points.
+ 8.5 x 11 inch transparency with a straight line etched across the middle.
+ 153 graduate students and post docs in Introductory Biostatistics.
+ Latin square.
+ 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).

University of Nebraska Lincoln

'You Draw It' Feature

(New York Times, 2015)



(Katz, 2017)

Readers are asked to input their own assumptions about various metrics and compare how these assumptions relate to reality.

- Family Income affects college chances (Aisch, Cox, and Quealy, 2015)
- Just How Bad Is the Drug Overdose Epidemic? (Katz, 2017)
- What Got Better or Worse During Obama’s Presidency (Buchanan, Park, and Pearce, 2017)

In 2015, the New York Times developed a You Draw it feature where readers are asked to input their own assumptions about various metrics and compare how these assumptions relate to reality.

The New York Times team utilizes **Data Driven Documents (D3)** that allows readers to predict these metrics through the use of drawing a line on their computer screen with their mouse.

Research Objectives

1. Validate 'You Draw It' as a method for graphical testing, comparing results to the less technological method utilized in Mosteller et al. (1981).
2. Extend the study with formal statistical analysis methods in order to better understand the perception of linear regression.

The two objectives of my current research are to:

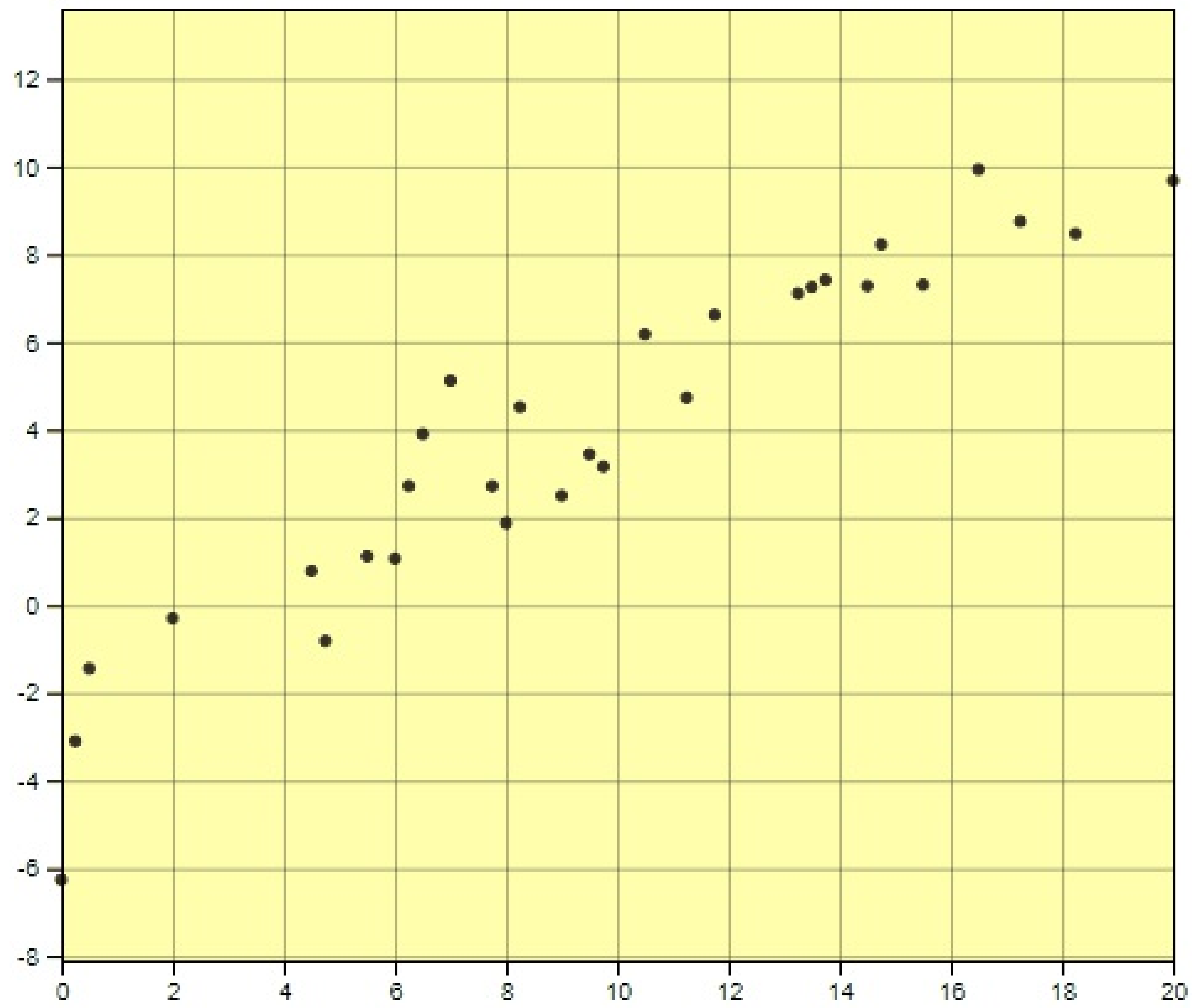
1. Validate 'You Draw It' as a method for graphical testing, comparing results to the less technological method utilized in Mosteller et al. (1981).

2. Extend the study with formal statistical analysis methods in order to better understand the perception of linear regression.

University of Nebraska Lincoln

'You Draw It' Task

Study Participant Prompt: *Use your mouse to fill in the trend in the yellow box region.*



Here we see an example of a "You Draw It" task plot used in the study. Participants are prompted to "Use your mouse to fill in the trend in the yellow box region. The yellow box region moves along as the participant draws their trend-line until the yellow region disappears."

Task plots were created using Data Driven Documents (D3), a JavaScript-based graphing framework that facilitates user interaction. We then integrate this into RShiny using the r2d3 package.

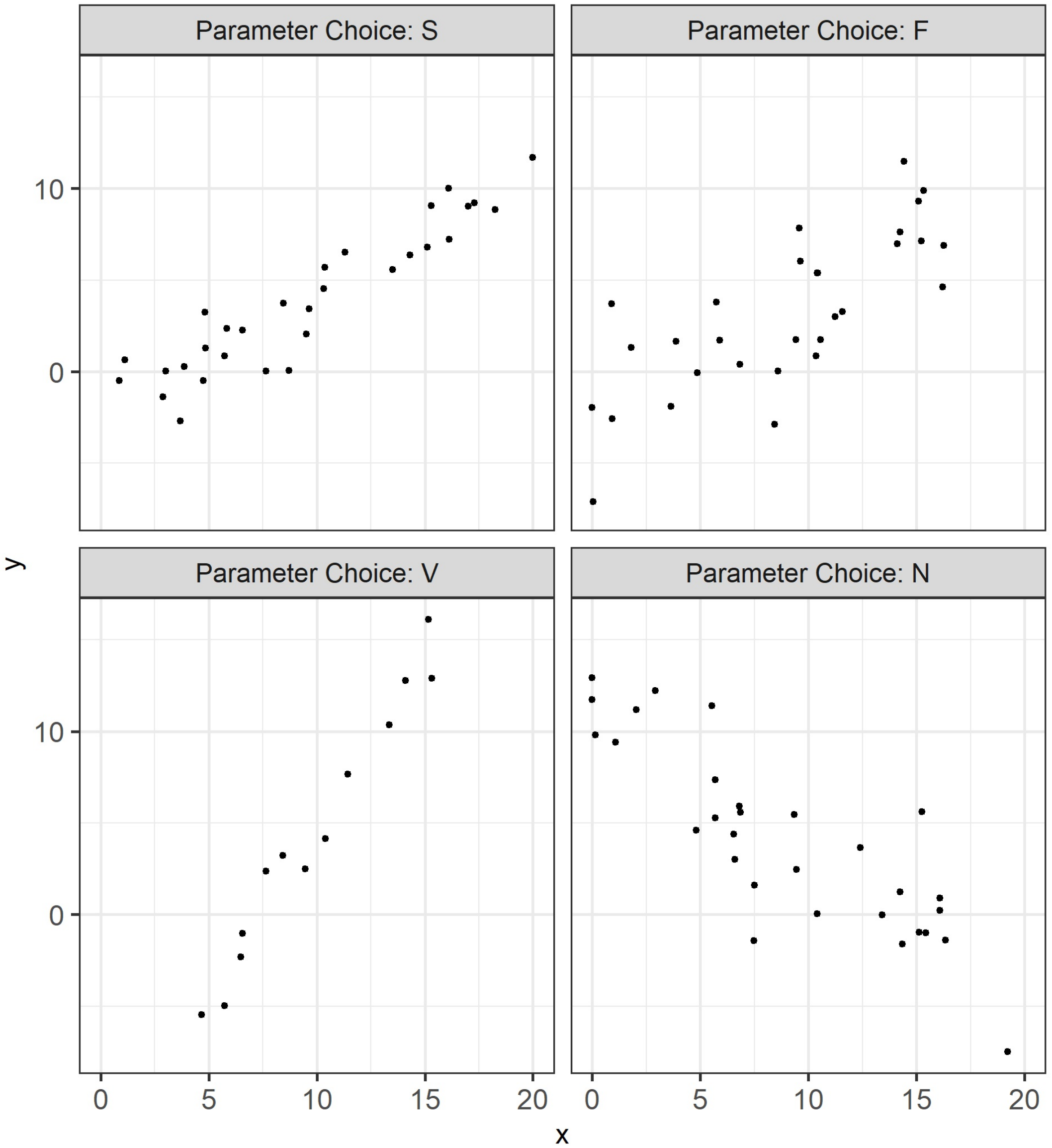
Data Generation

\(N = 30\) points \((x_i, y_i), i = 1,...N\) were generated for \(x_i \in [x_{min}, x_{max}]\).

Data were simulated based on linear model with additive errors: \begin{equation} y_i = \beta_0 + \beta_1 x_i + e_i \end{equation}

where \(e_i \sim N(0, \sigma^2).\)

Parameters \(\beta_0\) and \(\beta_1\) were selected to reflect the four data sets used in Mosteller, Siegel, Trapido, et al. (1981).



Data were generated following a linear model with additive errors.

Model equation parameters, β_0 and β_1, were selected to reflect the four data sets (F, N, S, and V) used in Mosteller et al. (1981).

+ **S:** positive slope; small variance; $x \in [0, 20]$.
+ **F:** positive slope; a large variance; $x \in [0, 20]$.
+ **V:** steep positive slope; small variance; $x \in [4, 16]$.
+ **N:** negative slope; large variance; $x \in [0, 20]$.

Study Design

- Participants recruited through Twitter, Reddit, and direct email in May 2021.
- A total of 35 individuals completed 119 unique you draw it task plots.
- Data sets were generated randomly, independently for each participant at the start of the experiment.
- Participants shown 2 practice plots followed by 4 task plots randomly assigned for each individual in a completely randomized design.
- Experiment conducted and distributed through an RShiny application found **here**.

Participants were recruited through through Twitter, Reddit, and direct email in May 2021. The experiment was conducted and distributed through an RShiny application. Participants were first shown 2 practice plots followed by the 4 You Draw It task plots randomly assigned for each individual in a completely randomized design.

Model Data

For each participant, the final data set used for analysis contains:

- \(x_{ijk}\), \(y_{ijk,drawn}\), \(\hat y_{ijk,OLS}\), \(\hat y_{ijk,PCA}\)

for

- parameter choice \(i = 1,2,3,4\),
- participant j = \(1,...N_{participant}\)
- \(x_{ijk}\) value corresponding to increment \(k = 1, ...,4 x_{max} + 1\).

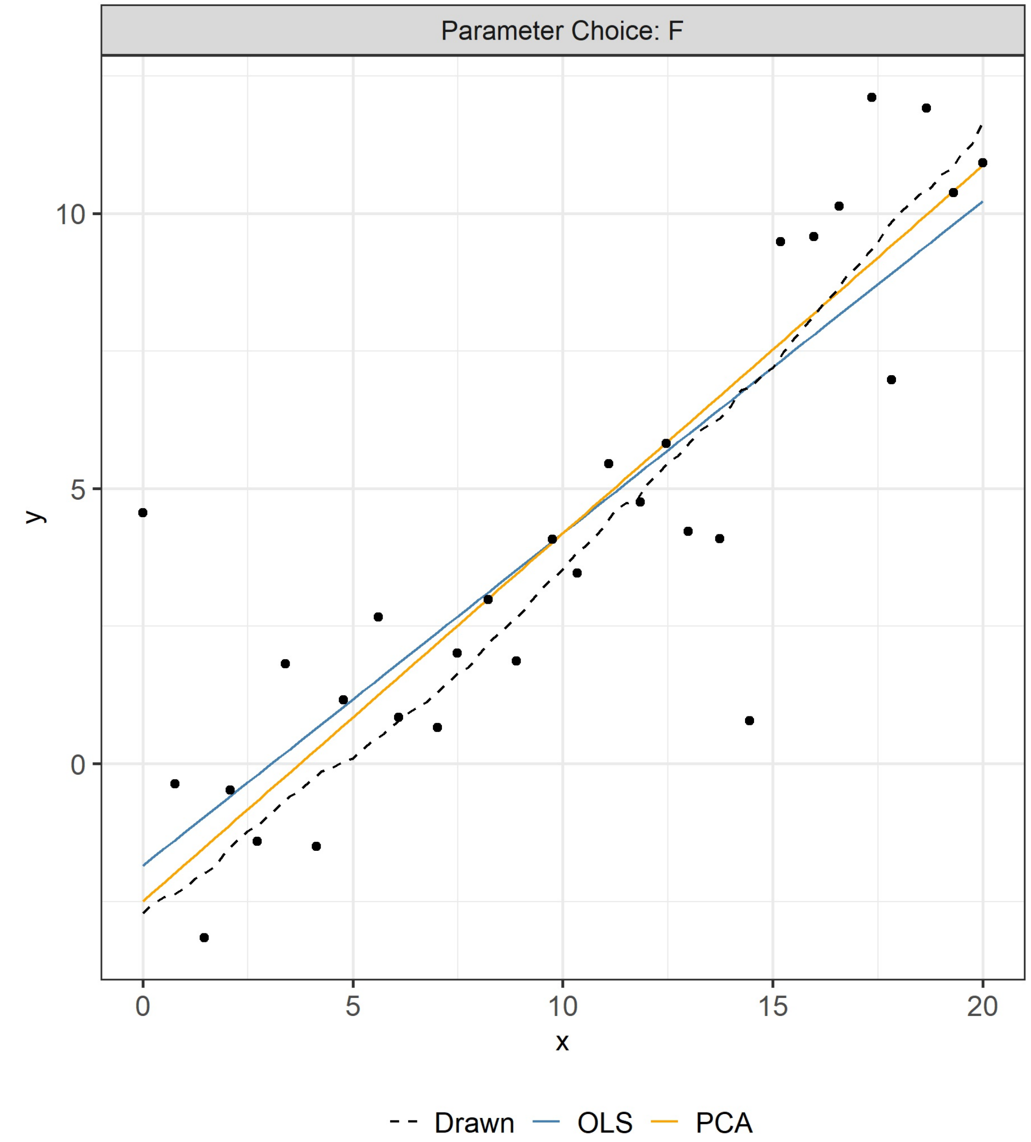
Vertical residuals between the drawn and fitted values were calculated as:

- \(e_{ijk,OLS} = y_{ijk,drawn} - \hat y_{ijk,OLS}\)
- \(e_{ijk,PCA} = y_{ijk,drawn} - \hat y_{ijk,PCA}\).

We compare the participant drawn line to two regression lines determined by ordinary least squares regression and regression based on the principal axis. The figure illustrates the difference between an OLS regression line which minimizes the vertical distance of points from the line and a regression line based on the principal axis (Principal Component) which minimizes the Euclidean distance of points (orthogonal) from the line.

Here we see an example of the feedback data from one you draw it plot. For 0.25 increments across the domain, we have the participant drawn values, the fitted values from the ordinary least squares regression, and the fitted values from the regression based on the principal axis.

We are mainly interested in the deviation of the participant drawn line from the fitted regression lines. So while it seems counter-intuitive, the residual actually becomes our response in this case.



Linear Trend Constraint

The **Linear Mixed Model** equation for each fit (OLS and PCA) residuals is given by:

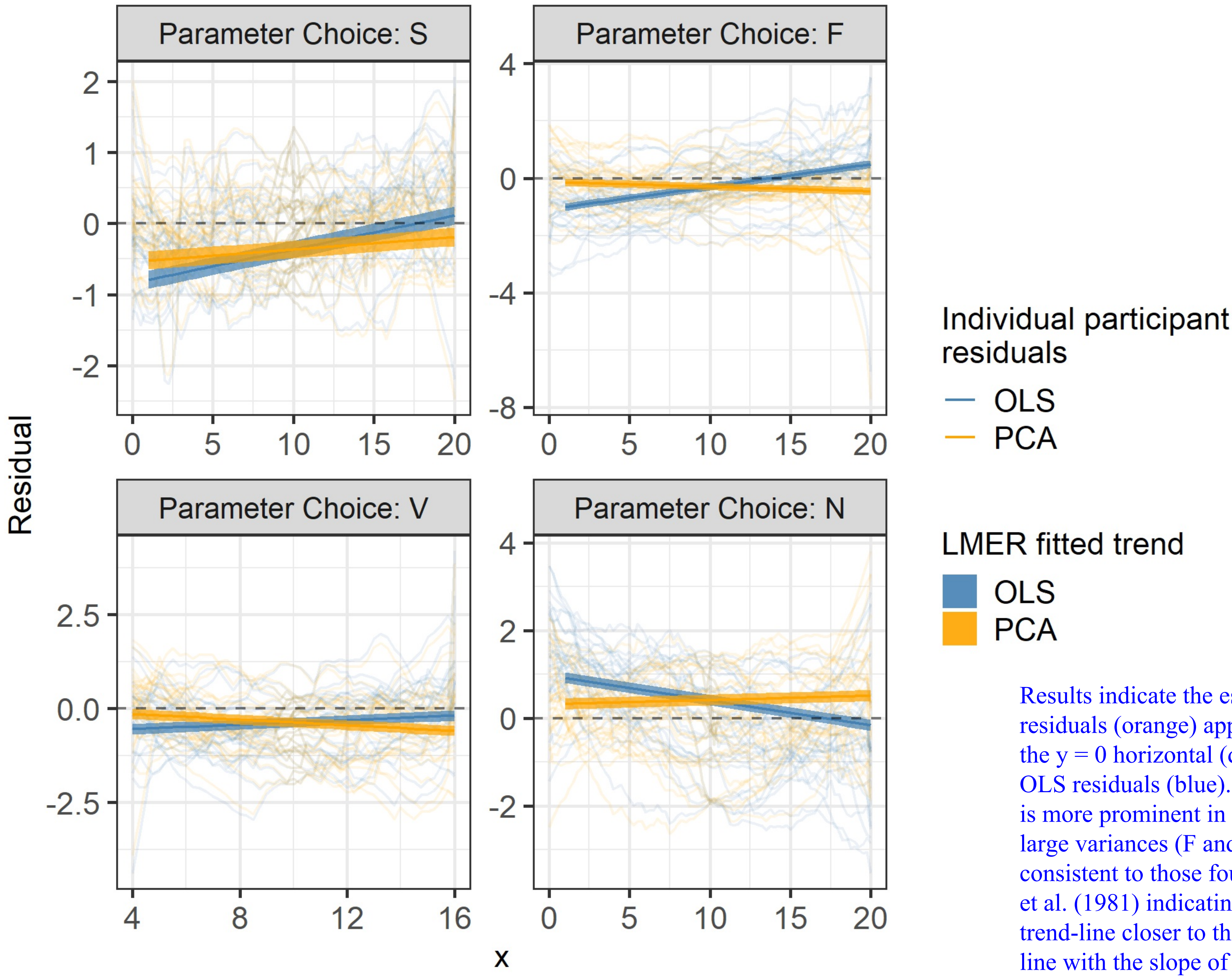
$$
\begin{equation} e_{ijk,fit} = \left[\gamma_0 + \alpha_i\right] + \left[\gamma_{1} x_{ijk} + \gamma_{2i} x_{ijk}\right] + p_{j} + \epsilon_{ijk} \end{equation}
$$

where

- $e_{ijk,fit}$ is the residual between the drawn and fitted y-values for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x-value corresponding to either the OLS or PCA fit
- γ_0 is the overall intercept
- α_i is the effect of the i^{th} parameter choice (F, S, V, N) on the intercept
- γ_1 is the overall slope for x
- γ_{2i} is the effect of the parameter choice on the slope
- x_{ijk} is the x-value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment
- $p_{j} \sim N(0, \sigma^2_{participant})$ is the random error due to the j^{th} participant's characteristics
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ is the residual error.

Using the `lmer` function in the lme4 package, a linear mixed model (LMM) is fit separately to the OLS and PCA residuals, constraining the fit to a linear trend.

Linear Trend Constraint



Results indicate the estimated trends of PCA residuals (orange) appear to align closer to the y = 0 horizontal (dashed) line than the OLS residuals (blue). In particular, this trend is more prominent in parameter choices with large variances (F and N). These results are consistent to those found in Mosteller et al. (1981) indicating participants fit a trend-line closer to the estimated regression line with the slope of based on the first principal axis than the estimated OLS regression line.

Smoothing Spline Trend

The **Generalized Additive Mixed Model** equation for each fit (OLS and PCA) residuals is given by:

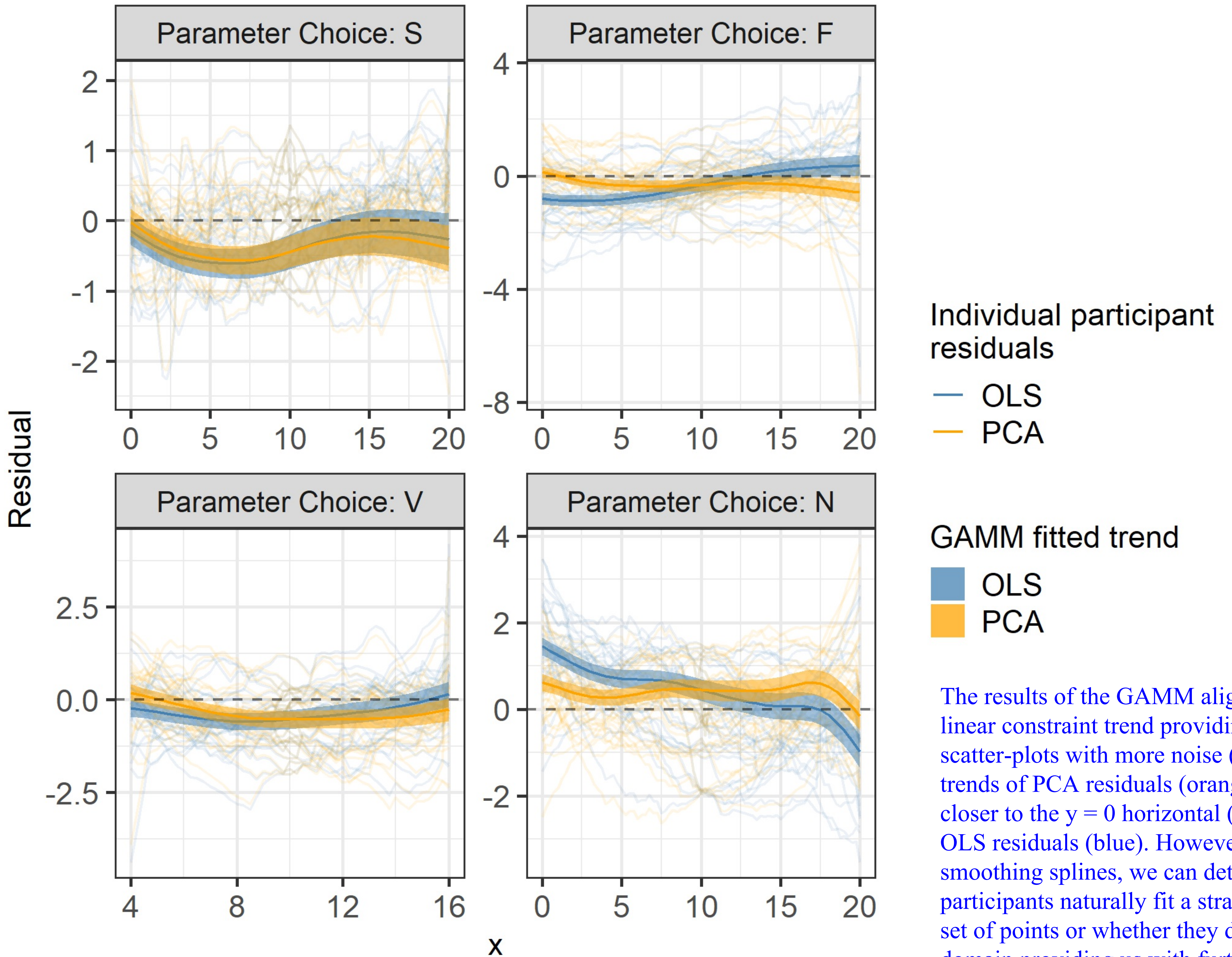
$$e_{ijk,fit} = \alpha_i + s_{i}(x_{ijk}) + p_{j} + s_{j}(x_{ijk})$$

where

- $e_{ijk,fit}$ is the residual between the drawn and fitted y-values for the i^{th} parameter choice, j^{th} participant, and k^{th} increment of x-value corresponding to either the OLS or PCA fit
- α_i is the intercept for the parameter choice i
- s_{i} is the smoothing spline for the i^{th} parameter choice
- x_{ijk} is the x-value for the i^{th} parameter choice, j^{th} participant, and k^{th} increment
- $p_{j} \sim N(0, \sigma^2_{participant})$ is the error due to participant variation
- s_{j} is the random smoothing spline for each participant.

Eliminating the linear trend constraint, the `bam` function in the mgcv package is used to fit a generalized additive mixed model (GAMM) separately to the OLS and PCA residuals to allow for estimation of smoothing splines.

Smoothing Spline Trend



The results of the GAMM align with those in the linear constraint trend providing support that for scatter-plots with more noise (F and N), estimated trends of PCA residuals (orange) appear to align closer to the y = 0 horizontal (dashed) line than the OLS residuals (blue). However, By fitting smoothing splines, we can determine whether participants naturally fit a straight trend-line to the set of points or whether they deviate throughout the domain providing us with further insight into the curvature humans perceive in a set of points.

Conclusion

Research Objectives:

1. Validate ‘You Draw It’ as a method for graphical testing, comparing results to the less technological method utilized in Mosteller et al. (1981).
2. Extend the study found in Mosteller et al. (1981) with formal statistical analysis methods for understanding the perception of linear regression.

Results:

- Estimated drawn trend-lines followed closer to the regression line based on the principal axes than the OLS regression line.
- Most prominent in data simulated with large variances.
- Humans perform “ensemble perception” in a statistical graphic setting.

The reproducibility of these results serve as validation of the 'You Draw It' tool and method.

1. Validate ‘You Draw It’ as a method for graphical testing, comparing results to the less technological method utilized in Mosteller et al. (1981).

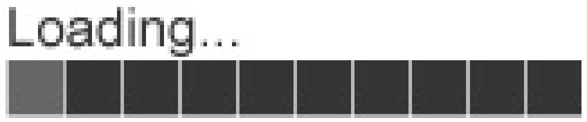
2. Extend the study found in Mosteller et al. (1981) with formal statistical analysis methods for understanding the perception of linear regression.

Results:
+ Estimated drawn trend-lines followed closer to the principal axes than the OLS regression line.
+ Most prominent in data simulated with large variances.
+ Humans perform “ensemble perception” in a statistical graphic setting as participants minimized the distance from the their regression line over both the x and y axis simultaneously

This study reinforces the differences between intuitive visual model fitting and statistical model fitting, providing information about human perception as it relates to the use of statistical graphics.

The reproducibility of these results serve as validation of the 'You Draw It' tool and method.

Future Work



- ✏ Implement the 'You Draw It' method in non-linear settings.
- Evaluate human ability to extrapolate data from trends.
- Use the tool to understand beliefs of real data such as climate change trends.
- Develop an R package designed for easy implementation of ‘You Draw It’ task plots.

Gif Source: photobucket.com

References

Aisch, G., N. Cohn, A. Cox, et al. (2016). Live Presidential Forecast. URL: https://www.nytimes.com/elections/2016/forecast/president.

Aisch, G., A. Cox, and K. Quealy (2015). You Draw It: How Family Income Predicts Children's College Chances. URL: https://www.nytimes.com/interactive/2015/05/28/upshot/you-draw-it-how-family-income-affects-childrens-college-chances.html.

Buchanan, L., H. Park, and A. Pearce (2017). You Draw It: What Got Better or Worse During Obama's Presidency. URL: https://www.nytimes.com/interactive/2017/01/15/us/politics/you-draw-obama-legacy.html.

Buja, A., D. Cook, H. Hofmann, et al. (2009). "Statistical inference for exploratory data analysis and model diagnostics". In: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367.1906, pp. 4361–4383.

Carpenter, P. A. and P. Shah (1998). "A model of the perceptual and conceptual processes in graph comprehension." In: Journal of Experimental Psychology: Applied 4.2, p. 75.

Chandar, N., D. Collier, and P. Miranti (2012). "Graph standardization and management accounting at AT&T during the 1920s". In: Accounting History 17.1, pp. 35–62.

Chong, S. C. and A. Treisman (2003). "Representation of statistical properties". In: Vision research 43.4, pp. 393–404.

— (2005). "Statistical processing: Computing the average size in perceptual groups". In: Vision research 45.7, pp. 891–900.

Ciccione, L. and S. Dehaene (2021). "Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots". In: Cognitive Psychology 128, p. 101406.

Cleveland, W. S. and R. McGill (1984). "Graphical perception: Theory, experimentation, and application to the development of graphical methods". In: Journal of the American statistical association 79.387, pp. 531–554.

References

Cleveland, W. S. and R. McGill (1985). "Graphical perception and graphical methods for analyzing scientific data". In: Science 229.4716, pp. 828–833.

Finney, D. (1951). "Subjective judgment in statistical analysis: An experimental study". In: Journal of the Royal Statistical Society: Series B (Methodological) 13.2, pp. 284–297.

Gouretski, V. and K. P. Koltermann (2007). "How much is the ocean really warming?" In: Geophysical Research Letters 34.1.

Green, T. M. and B. Fisher (2009). "The personal equation of complex individual cognition during visual interface interaction". In: Workshop on Human-Computer Interaction and Visualization. Springer. , pp. 38–57.

Harms, H. (1991). "August Friedrich Wilhelm Crome (1753-1833) Autor begehrter Wirtschaftskarten". In: Cartographica Helvetica 3, pp. 33–38.

Hofmann, H., L. Follett, M. Majumder, et al. (2012). "Graphical tests for power comparison of competing designs". In: IEEE Transactions on Visualization and Computer Graphics 18.12, pp. 2441–2448.

Katz, J. (2017). You Draw It: Just How Bad Is the Drug Overdose Epidemic? URL: https://www.nytimes.com/interactive/2017/04/14/upshot/drug-overdose-epidemic-you-draw-it.html.

Lewandowsky, S. and I. Spence (1989). "The perception of statistical graphs". In: Sociological Methods & Research 18.2-3, pp. 200–242.

Mosteller, F., A. F. Siegel, E. Trapido, et al. (1981). "Eye fitting straight lines". In: The American Statistician 35.3, pp. 150–152.

Playfair, W. (1801). "The statistical breviary; shewing, on a principle entirely new, the resources of every state and kingdom in Europe, Wallis, Londres". In: Press, Chicago.

References

Spence, I. (1990). "Visual psychophysics of simple graphical elements." In: Journal of Experimental Psychology: Human Perception and Performance 16.4, p. 683.

Unwin, A. (2020). "Why is data visualization important? what is important in data visualization?" In: Harvard Data Science Review 2.1.

Van Opstal, F., F. P. de Lange, and S. Dehaene (2011). "Rapid parallel semantic processing of numbers without awareness". In: Cognition 120.1, pp. 136–147.

Vanderplas, S., D. Cook, and H. Hofmann (2020). "Testing Statistical Charts: What makes a good graph?" In: Annual Review of Statistics and Its Application 7, pp. 61–88.

VanderPlas, S. and H. Hofmann (2015). "Spatial reasoning and data displays". In: IEEE Transactions on Visualization and Computer Graphics 22.1, pp. 459–468.

— (2017). "Clusters beat trend!? testing feature hierarchy in statistical graphics". In: Journal of Computational and Graphical Statistics 26.2, pp. 231–242.

Walker, F. A. (2013). Statistical atlas of the United States based on the results of the ninth census 1870 with contributions from many eminent men of science and several departments of the government.

Wickham, H. (2011). "ggplot2". In: Wiley Interdisciplinary Reviews: Computational Statistics 3.2, pp. 180–185.

Wilkinson, L. (2013). The grammar of graphics. Springer Science & Business Media.

Yates, J. (1985). "Graphs as a managerial tool: A case study of Du Pont's use of graphs in the early twentieth century". In: The Journal of Business Communication (1973) 22.1, pp. 5–33.

Thank you!

emily.robinson@huskers.unl.edu

www.emilyarobinson.com

earobinson95

University of Nebraska Lincoln