CHAPTER 1. SIGNS OF THE SINE ILLUSION – WHY WE NEED TO CARE

Accepted to the Journal of Computational and Graphical Statistics, July 2014

1.1 Introduction

Graphics are powerful tools for summarizing large or complex data, but they rely on the main premise that any graphical representation of the data has to be "true" to the data (see e.g. **???**). That is, a measurable quantity of a graphical element in the representation has to directly reflect some aspect of the underlying data. Generally, we see a lot of discussion on keeping true to the data in the framework of (ab)using three dimensional effects in graphics. **?** goes as far as defining a *lie-factor* of a chart as the ratio of the size of an effect in the data compared to the size of an effect shown, with the premise that any large deviations from one indicate a misuse of graphical techniques. Computational tools help us ensure technical trueness – but this brings up the additional question of how we deal with situations that involve innate inability or trigger learned misperceptions in the audience. In this paper we want to raise awareness for one of these situations, show that it occurs frequently in our dealings with graphics and provide a set of strategies for solving or avoiding it.

(a) Scatterplot of Ozone and Temperature in Houston, 2011. A loess fit is overlaid to show the overall trend.

(b) Scatterplots of Ozone and Temperature de-trended according to the loess fit in (a).

Figure 1.1: Scatterplots of Ozone and Temperature in Houston, 2011. The increase in variability over the temperature range is more pronounced in the de-trended plot on the right.

As a first example let us consider the relationship between ozone concentration and temperature. Ozone concentrations were measured from 21 locations in the Houston area (**?**), and temperature data is provided by the NCDC (**?**) site at Hobby International Airport, located near the center of Houston.

Figure **??** shows daily measurements of 8-hour average ozone concentration and temperature at several sites in Houston, for days in 2011 with temperatures above 45°F and dew points of less than 60°F. A loess smooth line is added for reference. These types of plots are often used to give an overview of the relationship between two variables. The trendline summarizes this relationship, while the points show raw measurement to allow an assessment of the overall size of the data, the amount of (marginal) variability presented, as well as the (conditional) variability along the trendline. It is the latter task that we cannot satisfactorily complete. While we might agree that there is an increase in variability of ozone concentrations for temperatures above 80°F, we will not doubt homogeneity elsewhere based on figure **??**.

This evaluation changes when considering figure **??**: the scatterplot shows a loess based detrended residual of temperature. A previously almost invisible increase in variability of ozone measurements with increasing temperatures now becomes apparent.

This phenomenon, caused by the change in the slope of the trend line, is known as the *sine illusion* in the literature on cognition and human perception or *line width illusion* in the statistical graphics literature.

The illusion is a frequent occurrence in statistical graphics, and displays should therefore be thoughtfully considered to minimize its effect visually and acknowledge its influence. In the cognitive literature, **?** first documented the illusion in the context of vertical lines along a sinusoidal curve. Figure **??** shows a sketch of this: line segments are centered evenly spaced along the curve. Line segments are of equal length but appear longer in the peaks and troughs due to the illusion. The parameters that influence the strength of the illusion are the amplitude of the curve and the length of the line segments. As the length of the line segments increases, the apparent difference in the length of the line segments decreases. Any modification that increases the change in slope under which the curve appears, such as an increase in the amplitude of the curve or a more extreme aspect ratio, reinforces the apparent difference in line lengths.

More recently the illusion has been shown in non-sinusoidal curves (**????**), but the underlying effect seems to be the same, in the sense that the illusion is not triggered by the periodic

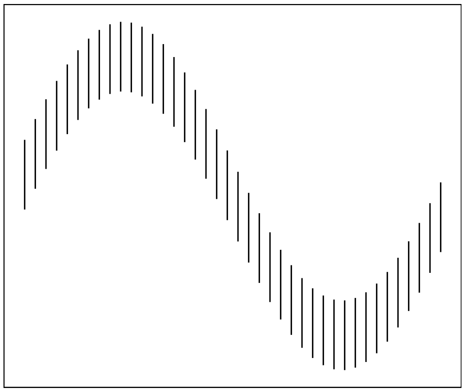
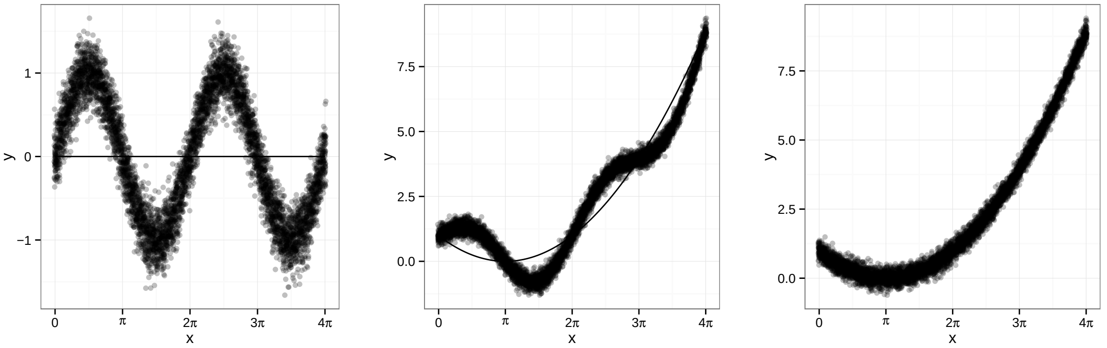


Figure 1.2: The original sine illusion, demonstrated on evenly spaced vertical lines centered around a sinusoidal curve of $f(x) = \sin(x)$. The lines in the peak and trough of the curve appear to be longer than in the other regions.

nature of the underlying trendline but only by changes to its slope. Figure **??** shows three panels, which all exhibit the illusion. From left to right, the trend stems from (a) a periodic function, (b) a periodic component added to an exponential function, and (c) an exponential function on its own. While all three graphs seem to show nonconstant variance along the main trend; in reality, it is constant. Clearly, the illusion does not rely on the periodicity of the function for which it was named, but is a symptom of the change in curvature that comes with the periodicity.



(a) Seasonality, No Trend (b) Seasonality and Trend (c) Trend, No Seasonality

Figure 1.3: Set of three scatterplots of simulated data with constant variance. Plot (a) shows seasonality without any underlying trend, (b) shows seasonality superimposed on a quadratic trend, and (c) shows a quadratic trend without seasonality. Though all three sets of simulated data have constant variance, none of the variances appear constant due to the sine illusion.

Next, we give an overview of the perceptual and statistical literature regarding this illusion.

1.1.1 The Sine Illusion in Statistical Graphics

The sine illusion demonstrated in figures **??** and **??** has been frequently noted in statistical graphics, though usually not as an optical illusion. Rather, the problem is typically identified as the difficulty of visually subtracting two curves, and the resulting erroneous conclusions when this process goes awry. Figure **??** presents the possibly oldest example of this common phenomenon (**??**): Playfair's chart of the balance of trade between England and the East Indies shows time series of the trade value for imports and exports between the countries in the 18th century. The shaded area on the chart is named "balance against England", suggesting that the difference between the lines is of main importance. This difference in trading values is encoded as the difference between the lines along the vertical axis. However, the vertical distance between two lines provides a much less visually salient cue than the orthogonal width between the lines. This results in an underestimation (**?**) of the difference in trades around 1763, which is of a much higher (about 1.5 fold) magnitude as around 1770, but appears much smaller. In more modern visualizations, bivariate area charts and "stream graphs" (**?**) commonly produce the illusion (see an example at http://bl.ocks.org/mbostock/3894205).

Figure 1.4: Playfair's graph of exports to and imports from the East Indies demonstrates that the line width illusion is not only found on sinusoidal curves but is present whenever the slope of the lines change dramatically. The increase in both imports and exports circa 1763 does not appear to portray as large of a deficit as that in 1710, even though they are of similar magnitude.

1.1.2 Perceptual Explanations for the Sine Illusion

While not thoroughly examined in the sensation and perception literature, the sine illusion has been classified as part of a group of geometrical optical misperceptions related to the Müller-Lyer illusion (**?**) or the Poggendorf illusion (**?**), which puts the illusion into the framework of context-based illusions. **?** suggest that the sine illusion occurs due to misapplication of perceptual experience with the three-dimensional world to a two-dimensional "artificial" display of data.

Experience with real-world objects suggests that the stimulus of figure ?? is very similar to a slightly angled top view of the 3-dimensional figure of a strip or ribbon describing waves in a third dimension, such as e.g. a road does on rolling hills. This is sketched out in figure ??. Our experience suggests immediately that changes in the width of the road are unlikely and resolves the illusion. While figure ?? shows the line segments slightly angled towards each other, figure ?? shows a variation of the same plot with a vanishing point set further away from the viewer. This makes the line segments almost parallel to each other and therefore more closely resembles the sketch of figure ??, in which the sine illusion was originally presented.

(a) Perspective plot of sine illusion

(b) Perspective plot, vanishing point near infinity.

Figure 1.5: Two different perspective projections of the same data responsible for the sine illusion. The first projection angles the lines and appears more natural, but the second projection suggests that the lines do not need to be angled to create the same three-dimensional impression.

Recreating the three-dimensional context of the sine illusion might resolve the distortion, even if increasing the dimensionality of a graph is generally not recommended (??) (though ? suggests that in certain cases additional dimensions are not misleading). While creating a three-dimensional projection of two-dimensional data might counteract the illusion, the process of projecting the data accurately into a higher dimension is not simple. The projection that best resolves the illusion likely is highly subjective and influenced by choices of angle and color gradient for depth cues. As there is not a single three-dimensional projection that corresponds to the two-dimensional data, this approach would only produce further visual ambiguity.

Further complicating the situation, the illusion itself is insidious – we trust our vision implicitly, to the point that when we understand something, we say “I see”. This trust in our visual perception is seldom called into question, for our perception is optimized for interaction with

Figure 1.6: The sine illusion with two individual lines highlighted. Horizontal grid lines do not help to resolve the illusion, even though they provide a clear basis for comparison of line lengths. Readers are much better at assessing the length of the two singled out line segments; they are equal.

a three-dimensional world. Artificial two-dimensional situations (such as graphs and pictures) may accurately represent the data and still produce a misleading perceptual experience.

The contextual cues of the overall trend are critical to the sine illusion's effect; the illusion only holds when a substantial portion of the graph is considered simultaneously, which triggers our innate ability of perceiving one whole rather than the individual parts it consists of (principle of grouping; **?**). Considering only two line segments at a time resolves the illusion. The bold lines in figure **??** are clearly of the same length. Comparisons of individual line lengths is visually a fairly simple task, and is done with a relatively high accuracy (**?**). **?** contains a more thorough discussion of how much surrounding context is required for the illusion to persist.

1.1.3 Geometry of the Illusion

In figure **??** we have seen that the our preference in evaluating line width is to assess *orthogonal* width rather than the difference along the vertical axis. Figure **??** demonstrates the change in orthogonal width as the slope of the line tangent to the graph of f changes; these changes correspond to our perception of apparent line length.

Figure 1.7: The sine illusion with lines orthogonal to the tangent line at $f(x)$. The perception that the vertical length changes with $f(x)$ corresponds to changes in actual orthogonal width due to the change in the visual (plotted) secant angle. The strength of the perceptual effect depends in part on the aspect ratio of the graph, as shown in the second image, which has an aspect ratio of 2 compared to the first figure's aspect ratio of 1. This correspondingly multiplies the strength of the effect by 2.

The illusion is most pronounced in regions where the angle between the orthogonal and the vertical line is large. Changes to the aspect ratio therefore have a major impact on the strength of the sine illusion. Any change that alleviates the difference between perceived width and the perpendicular width, such as banking to 45° (**?**), will alleviate the effect but not completely overcome it. The perceived length of the vertical line changes with the angle of the line perpendicular to the slope of $\sin(x)$, suggesting that the sine illusion stems from a conflict between the visual system's perception of figure width and the mathematical judgement necessary to determine the length of the vertical lines.

Our preference for assessing figure width based on the orthogonal width suggests that the underlying illusion may be a function of geometry rather than some unknown visual or neural process that occurs subconsciously. In this case it may be possible to correct the graphical display for the illusion to minimize its misleading effect. A geometrical correction that –at least temporarily– counteracts the illusion would be a valuable tool in visual analysis, as this illusion very persistently affects our judgment of very common tasks such as e.g. the assessment of conditional variability of data along a trend line.

Simply raising people's awareness of the presence of this illusion is not enough, as it is incredibly difficult, if not impossible, to overcome this illusion even when we are aware of its presence: our brains simply cannot "un-see" it.

What follows is a compilation of several approaches to correct for or mitigate the effect of the illusion. Our primary intention here is to demonstrate the persaviness of the illusion is and the extreme measures necessary to remove its effect.

1.2 Breaking the Illusion

The sine illusion is caused by a conflict between vertical width, which is the width that we want onlookers to assess visually, and orthogonal width, which is the width that the onlooker perceives. This difference can be expressed as a function in the slope of the underlying trend line. This provides the basis for adjusting the vertical width for the perceived orthogonal width.

We consider the following three approaches:

1. separating the trend and the variability,
2. transformation of x: adjusting the slope to be constant by reparametrizing the x axis, and
3. transformation of y: adjusting y values to make conditional variability appear correctly by adjusting accordingto orthogonal width.

Each of these ideas is discussed in more detail in this section.

(a) Data

(b) Trend and Residuals

Figure 1.8: Describe the conditional variability of the points along the x axis in (a). Is your description consistent with the residual plot in (b)?

1.2.1 Trend Removal

?? discuss the perceptual difficulty of judging the difference between two curves plotted in the same chart, and alternatively, recommend to display the difference between the two curves directly. This is in line with recommendations for good graphics to 'show the data' rather than make the reader derive some aspect of it (e.g. **?**). In particular, de-trending data to focus on residual structure is the generally accepted procedure for assessing model fit. Figure **??**(a) shows a scatterplot of data with a trend. A loess smooth is used to estimate the trendline. A visual assessment of variability along this trendline might result in a description such as 'homogeneous variance or slightly increasing variance for negative x, followed by a dramatic decrease in vertical variability for positive x'. Once the residuals are separated from the trendline as shown on the right hand side of the figure, it becomes apparent that this first assessment of conditional variability was not correct, and the decreasing variance along the horizontal axis becomes visible.

While the illusion is not apparent when trend line and variability in the residual structure are shown separately, the separation makes it more difficult to evaluate the overall pattern in the data, as we must base any judgment on two charts; either by combining information from two graphs or by mentally re-composing the original graph (at which point, the sine illusion becomes a factor). To minimize cognitive demands we ideally want to tell the whole story with a single graph, in particular because in many situations we may not be able to show multiple graphs.

Additionally, removing the trend requires an initial model, making any plots produced using that fit conditional on the assumptions necessary to obtain that model fit. In many situations, this may be undesireable. In particular, we typically view the data before fitting even a rudimentary model, and the sine illusion may influence even these initial modeling decisions.

1.2.2 Transformation of the X-Axis

As the sine illusion is driven by changes in the slope of trends between variables, we can counteract the illusion by removing these changes, transforming the x axis such that the absolute value of the slope is constant and forcing the corresponding orthogonal width to represent the conditional variability. In order to describe this transformation of the x axis mathematically, let us assume that the relationship between variables X and Y is given by a model of the form

$$y = f(x) + \varepsilon,$$

where f is some underlying function (either previously known or based on a model fit). Further let us assume that f is differentiable over the region of observed data.

Ideally, the correction would force all lines to appear under the same slope, i.e. we want to find a transformation $T(x)$ of x, such that $f(T(x))$ is a piece-wise linear function, where each piece has the same absolute slope. This transformation has an effect similar to "banking to 45°" in a piecewise manner.

Let a and b be the minimum and maximum of the x-range under consideration. Then for any value $x \in (a, b)$ the following transformation results in a function with constant absolute slope:

$$(f \circ T)(x) = a + (b-a)\left(\int_a^x |f'(z)| dz\right) / \left(\int_a^b |f'(z)| dz\right), \tag{1.1}$$

1.2.2.1 Derivation of the X Transformation

As the slope is determined by the aspect ratio, we are free to choose it and w.l.o.g. we get for each piece T_i:

$$f(T_i(x)) = \pm ax + b_i.$$

This means that T_i is essentially an inverse of function f, with each piece defined by the intervals on which the inverse of f exists: let $\{x_0 = \min(x), x_1, ..., x_{K-1}, x_K = \max(x)\}$ be the set of values with local extrema enhanced by the boundaries of the x-range, i.e. $f'(x_i) = 0$ for $i = 1, ..., K-1$ and $f'(x) \neq 0$ for any other values of x. Then each interval of the form (x_{i-1}, x_i) defines one piece T_i of the transformation function $T(x)$. We will define T_i now as a combination of a linear scaling function and the inverse of f, which we know exists for interval (x_{i-1}, x_i).

Let function $s = {}_{[a,b]}s^{[c,d]}$ be the linear scaling function that maps the interval (a, b) linearly to the interval (c, d). This function is formally defined as

$$s(x) = {}_{[a,b]}s^{[c,d]}(x) = (x - a)/(b - a) \cdot (d - c) + c \text{ for all } x \in (a, b).$$

Note that the slope of function s is given as

$$s'(x) = (d - c)/(b - a).$$

Two scaling functions can be evaluated one after the other, only if the image (i.e. y-range) of the first coincides with the domain (i.e. x-range) of the second. This consecutive execution results in another linear scaling:

$${}_{[e,f]}s^{[c,d]}\left({}_{[a,b]}s^{[e,f]}(x)\right) = {}_{[a,b]}s^{[c,d]}(x)$$

In our situation let the scaling function s be given as:

$${}_{[c,d]}s^{f([x_{i-1},x_i])}(x) = f(x_{i-1}) + (x - c)/(d - c) \cdot (f(x_i) - f(x_{i-1}))),$$

where $f([x_{i-1}, x_i])$ is defined as the interval given by $(\min(f(x_{i-1}), f(x_i)), \max(f(x_{i-1}), f(x_i)))$. Note that s has either a positive or negative slope depending on whether $f(x_{i-1})$ is smaller or larger than $f(x_i)$, respectively.

Then the transformation in the x-axis, $T(x)$ is defined piecewise as a combination of T_i, where each T_i is given as:

$$T_i(x) \quad = \quad f^{-1}\left({}_{[c_i,d_i]}s^{f([x_{i-1},x_i])}(x)\right). \tag{1.2}$$

Using this definition for the transformation makes $f(T(x))$ a piece-wise linear function with parameters c_i and d_i, i.e. for $x \in (c_i, d_i)$ we have

$$f(T(x)) = f(f^{-1}({}_{[c_i,d_i]}s^{f([x_{i-1},x_i])}(x))) = {}_{[c_i,d_i]}s^{f([x_{i-1},x_i])}(x).$$

Correspondingly, the slope of $f(T_i(x))$ is $(f(x_i) - f(x_{i-1})))/(d_i - c_i)$. In order to make the slope the same on all pieces T_i of T, we need to define c_i and d_i with respect to the function values on the interval (x_{i-1}, x_i). There are various options, depending on how closely the x-range of T should reflect the original range: for $[c_i, d_i] = \text{range}\,(f([x_{i-1}, x_i]))$ the new x-range is the range of f on (x_{i-1}, x_i), but with the advantage that the scaling function simplifies to the identity or a simple shift.

In order to preserve the original x-range, we need to invest into a bit more work for the scaling. With an identity scaling, each T_i maps from the range of f on (x_{i-1}, x_i) to the same range. Overall we can therefore set up the function T to map from the interval given by the sum of the function's 'ups' and 'downs', i.e. $(0, \sum_{i=0}^{K} |f(x_i) - f(x_{i-1})|)$, to the range of f on (x_0, x_K). This ensures that all pieces $f(T_i)$ have the same slope (of $|1|$).We can then use another - global - linear scaling function to map from the range of x, i.e. interval (x_0, x_K) to $(0, \sum_{i=0}^{K} |f(x_i) - f(x_{i-1})|)$, yielding a transformation function T of

$$T(x) = (f^{-1} \circ {}_{[c_i,d_i]}s^{f([x_{i-1},x_i])} \circ {}_{(x_0,x_K)}s^{(0,\sum_{i=0}^{K} |f(x_i)-f(x_{i-1})|)})(x),$$

where c_i and d_i are given as

$$c_i = \sum_{j=0}^{i-1} |f(x_j) - f(x_{j-1})| \text{ and } d_i = \sum_{j=0}^{i} |f(x_j) - f(x_{j-1})|.$$

We can write the difference $|f(x_j) - f(x_{j-1})|$ as $\int_{x_{j-1}}^{x_j} |f'(z)| dz$. This shows equation (**??**).

1.2.2.2 Weighting the X Transformation

As the sine illusion depends on changing slope in the overall trend, re-parametrizing the x-axis in terms of the slope will make the data appear under a constant slope, thereby removing the effect of the illusion, while the transformed x-axis is changed from a linear representation of the x values to a 'warped' axis that continuously changes the scale of x to compensate for the changes in the slope. To emphasize this change in scale along the x axis, dots are drawn at the bottom of the chart to show the transformation's effect on equally spaced points along the x-axis.

Results from this transformation are demonstrated in Figure **??**.

While the transformation in equation (**??**) effectively removes the appearance of changing line lengths, we can see in practice that the illusion can be broken by a much less severe

(a) X axis transformation based on eqn. (**??**), corresponding to weighting of $w = 0$.

(b) Weighted Transformation, $w = 1/2$ (based on eqn. (**??**))

(c) Weighted Transformation, $w = 1/3$

(d) Weighted Transformation, $w = 1/4$

Figure 1.9: Examples of X axis transformations in the sine curve. Dots at the bottom of the graph show the transformation's effect on equally spaced points along the x-axis. Different amounts of weighting w correspond to differently strong corrections. In (a), x-spacing of the lines changes the extant width such that the absolute value of the slope is uniform across the whole range of the x axis resulting in the largest amount of correction. (b) - (d) reduce the correction in (a) towards successively more uniform spacings in x while still breaking the effects of the illusion.

transformation of the x axis. For that we introduce a shrinkage factor $w \in (0, 1)$ that allows a weighted approach in counteracting the illusion as:

$$(f \circ T_w)(x) = (1 - w) \cdot x + w \cdot (f \circ T)(x) \tag{1.3}$$

Note that for $w = 1$ the x-transformation is completely warped, while smaller values of w indicate a less severe adjustment against the sine illusion. Under weaker transformations, the data more closely reflect the original function $f(x)$. Figures **??** - **??** show the effect of different shrinkage coefficients w. As w decreases, the lines become more evenly spaced and the illusion begins to return.

The extent to which we can shrink the adjustment back to the original function varies with the aspect ratio of the chart and the shape of the function. It might also be influenced by the audience's experience with the sine illusion, resulting in very subjective choices of an “optimal weighting” for specific situations which minimizes distortion and maximizes the correspondence between inferences made from the data and inferences made using the visual display.

Note, that we only make use of the transformation T in the form of $f \circ T$. This allows us to avoid an explicit calculation of the transformation T, which in particular involves a computation of the inverse of f leading to potentially computation-intense solutions.

1.2.2.3 X Transformation Demonstration

In the example of the Ozone data shown in figure **??**, we can base a transformation of the x-axis on a loess fit of ozone concentration in daily temperature. Loess is particularly convenient for this transformation, as it enforces continuity conditions including differentiability of the fitted function; software allows us to obtain fits of both the function values and their derivatives.

Figure **??** shows the original data side-by-side with the transformed x-axis, demonstrating not only the effect of transformation of the x-axis, but also that the transformation is not overly misleading in this example. The granularity of the data in this example provides an implicit measure of the strength of the transformation along the x-axis and the transformation is also clearly evident in the labels along the x-axis.

(a) Original Data

(b) Transformed X Axis

Figure 1.10: Original data and data after x-transformation. The increasing variance is easier to see when x has been transformed, because the slope is now uniform.

1.2.3 Transformation in Y

Understanding the geometry of the sine illusion leads to another approach to counteracting the conflict between the orthogonal width and the vertical length of the segment.

Let again the function f describe the general relationship between variables X and Y.

As sketched out in figure **??** we want to first find the orthogonal (extant) width in a point $(x_0, f(x_o))$ on the graph, which corresponds to the perceived width, and then correct the vertical width accordingly to match with the audience's expectation.

The orthogonal width (see sketch in figure **??**) is given as the line segment between endpoints $(x_1, f_1(x_1))$ and $(x_2, f_2(x_2))$, where f_1 and f_2 denote the vertical shifts of function f by $-\ell/2$ and $\ell/2$, respectively, where ℓ is defined as the overall line length, $\ell > 0, \ell \in \mathbb{R}$. These endpoints are determined as the intersection of the line orthogonal to the tangent line in $(x, f(x))$ and the graphs resulting from the vertical shifts of f.

The function describing the orthogonal line through $(x_o, f(x_o))$ is given in point-vector form

Figure 1.11: General correction approach. This approach may require numerical optimization to obtain exact solutions for (x_1, y_1) and (x_2, y_2).

as

$$\begin{pmatrix} x_o \\ f(x_o) \end{pmatrix} + \lambda \begin{pmatrix} f'(x_o) \\ 1 \end{pmatrix},$$

for any real-valued λ. The advantage of using point vector form is that it allows us to solve for parameter λ easily, which gives us easy access to the extant (half-)widths, as:

$$|\lambda|\sqrt{1 + f'(x_o)^2}. \tag{1.4}$$

Eqn. (**??**) describes the quantity that we perceive rather than the quantity that we want to display ($\ell/2$), which leads us to a general expression of the correction factor as

$$\ell/2 \cdot \left(|\lambda|\sqrt{1 + f'(x_o)^2}\right)^{-1}.$$

Note that this yields in general two solutions: one for positive, one for negative values of λ corresponding to upper and lower (half-)extant width.

In order to get actual numeric values for λ, we need to find end points of the extant line width as solutions of intersecting the orthogonal line and the graphs of f_1 and f_2. We find these end points as solutions in x and λ of the system of equations:

$$x - x_o \quad = \quad \lambda f'(x_o) \tag{1.5}$$

$$f(x) - f(x_o) \quad = \quad -\lambda \pm \ell/2 \tag{1.6}$$

Note that the above system of equations involves function values $f(x)$, which implies that solving this system requires numerical optimization for any but the most simple functions f.

In the following two sections we make use of Taylor approximations of first and second order to find approximate solutions to end points as sketched out in figure **??**.

(a) Linear Approximation

(b) Quadratic Approximation

Figure 1.12: (a) uses a first-order Taylor series approximation to $f(x)$ and (b) uses a second-order Taylor series approximation to $f(x)$. The intersection of the function $f(x) \pm \ell/2$ and the orthogonal line, $(x_1, y_1), (x_2, y_2)$ must be obtained to determine the necessary correction factor.

1.2.3.1 Linear Approximation to $f(x)$

For the linear approximation we make use of $f(x) \approx f(x_0) + (x - x_0)f'(x_0)$, which together with equations **??** and **??** yields a correction factor in x_0 of

$$\ell_{\text{new}}(x_0) = \ell_{\text{old}}\sqrt{1 + f'(x_0)^2}.$$

Note that the linear method gives the same result as a varying slope extension from a trigonometric approach suggested by **?** and used in **?**.

A second-order Taylor polynomial approximation to $f(x)$ additionally accounts for the asymmetry in the extant widths on either side of the center trendline.

1.2.3.2 Quadratic Approximation to $f(x)$

Using the approximation $f(x) \approx f(x_0) + f'(x_0)(x - x_0) + 1/2f''(x_0)(x - x_0)^2$, the system of equations **??** and **??** simplifies to the following quadratic equation in λ:

$$f''(x_0)f'(x_0)^2\lambda^2 + 2(f'(x_0)^2 + 1)\lambda \pm \ell = 0,$$

which leads us to corrections for the half lengths as:

$$\begin{aligned} \ell_{\text{new}_1}(x_0) &= 1/2 \cdot \left(v + \sqrt{v^2 + f''(x_0)f'(x_0)^2 \cdot \ell_{\text{old}}}\right) \cdot v^{-1/2} & (1.7) \\ \ell_{\text{new}_2}(x_0) &= 1/2 \cdot \left(v + \sqrt{v^2 - f''(x_0)f'(x_0)^2 \cdot \ell_{\text{old}}}\right) \cdot v^{-1/2} & (1.8) \end{aligned}$$

where $v = 1 + f'(x_0)^2$.

1.2.3.3 Reformulation of the quadratic approximation

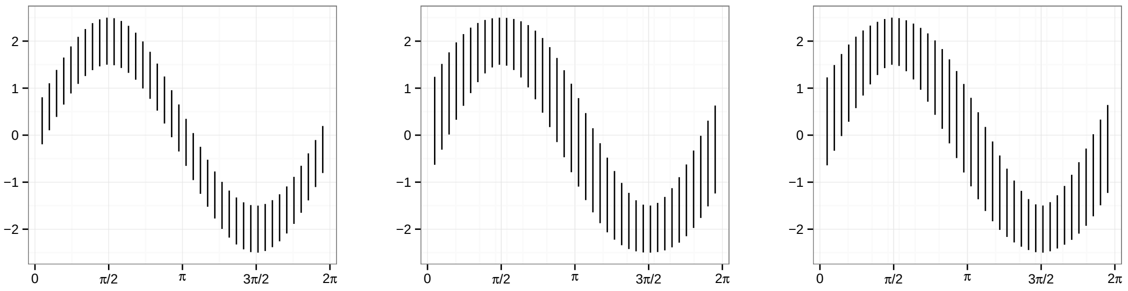
A quadratic equation in λ of the form

$$a\lambda^2 + b\lambda + c = 0, \tag{1.9}$$

where a, b, and c are real-valued parameters the solutions take on the form

$$\lambda_\pm = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \overset{*}{=} 2c\left(-b \pm \sqrt{b^2 - 4ac}\right)^{-1}.$$

* if $b \neq \pm \quad \sqrt{b^2 - 4ac}$, i. e. $a, c \neq 0$.



(a) Uncorrected (b) Linear correction (c) Quadratic correction

Figure 1.13: In the quadratic approximation top and bottom segments of the vertical lines are adjusted separately.

Application to quadratic approximation to f: in the example, we have the following equivalencies:

$$\begin{aligned} a &= f''(x_0)f'(x_0)^2 \\ b &= 2(1+f'(x_0)^2) \qquad\qquad > 0 \text{ for all } x \\ c &= \pm\ell \end{aligned}$$

For a valid solution for the correction factor, we have to assume that λ is a factor that extends the original extant width (in absolute value).

$$\lambda_{1/2} = \ell\left(v + \sqrt{v^2 \pm f''(x_0)f'(x_0)^2 \cdot \ell}\right)^{-1}$$

for $v = 1 + f'(x_0)$. This gives the results as shown in equations (**??**) and (**??**)

Adjusting the top and bottom segments of the vertical lines separately so that the extant width is constant breaks the illusion, but slightly distorts the sinusoidal shape of the peaks.

Figure **??** shows the correction factor based on a quadratic approximation compared to the untransformed data. Unlike the linear solution, the half-segments here are not necessarily of the same length, and thus there are separate correction factors for each half-segment.

1.2.3.4 Mathematical Properties of the Y Transformation

The quadratic correction breaks whenever the expression in the square root of eqn. (**??**) becomes negative, i.e. whenever $v^2 \pm \ell \cdot f''(x) \cdot f'(x)^2 < 0$. This happens for combinations of large values of ℓ, which signify a large vertical extent, or large conditional variability $E[Y|X]$, and simultaneous large changes in the slope of the main trend, i.e. large values of the curvature $f''(x)$. In the linear approximation of f the same situation leads to a massive overcorrection of the vertical lines, changing the shape of the 'corrected' function beyond recognition.

Similar to the correction of the x-axis, we can use a weighted approach to find a balance between counteracting the illusion and representating the original data:

$$\ell_{new_w}(x) = (1 - w) \cdot \ell_{old} + w \cdot \ell_{new}(x) \tag{1.10}$$

1.3 Transformations in Practice – a User Study

In order to more fully understand the sine illusion and test the proposed corrections, we created an applet to allow users to investigate the illusion's prominence with respect to its parameters. Users can examine the sine illusion by changing line length, the function's amplitude, and compare corrections in x-axis and y-values to uncorrected data. All corrections proposed in this paper are implemented in the applet located at http://glimmer.rstudio.com/srvanderplas/SineIllusion/.

We employed a second applet to collect data on users' preferences on the amount of correction used, i.e. we are interested in identifying a range of 'optimal weights' in each of the corrections. This applet presents users with a graph that is the result of a correction in x or y with a randomly selected starting weight value. Users are asked to adjust the graph until the illusion (a) is no longer apparent (adjustment of weights from the bottom) or (b) becomes visible (adjustments of weights from the top).

Both applets are implemented in `shiny` (**?**).

The graphs in the data-collection applet are adjusted using a plus/minus button to either increase or decrease the amount of correction used. Underlying this adjustment is the value of the weight w as defined in eqns. (**??**) and (**??**). The numerical value of w was hidden from the user to prevent anchoring to a specific numerical value.

A low initial weight (w_0 close to 0) indicates that the amount of correction is low and the response from a trial like this will give us an idea of the minimal amount of weight necessary

Figure 1.14: Screenshot of the shiny application used to collect information of observers' preference with respect to an optimal correction for the illusion under each of the transformations discussed in the previous section.

to break the illusion, while a high initial weight (w_0 close to 1) indicates that the data is fully corrected. We asked participants to change the amount of adjustment until the lines appear to be the same length assumes that the correction is overcorrecting in practice, and a response from this type of trial gives us an upper boundary for the amount of weighting preferred. Generally, responses from the two different types of trials do not result in the same threshold weight, but rather lead to a range of acceptable weights.

It is of additional interest to determine whether and how much these optimal weights are subject-specific or population-based, whether they depend on the initial weight, and how much within-subject variability we find compared to between-subject variability.

Figure **??** shows a screenshot of the applet used to collect user data. This applet is available online at http://glimmer.rstudio.com/srvanderplas/SineIllusionShiny/. Line length and function are controlled in this app, and we used the linear transformation for adjusting y values; the transformation does not break under any combination of parameters tested in this experiment.

We deployed the applet to participants recruited online, collecting their responses and other metadata. The results of the analysis suggest that the correction factors in X and Y are both preferrable to uncorrected data, but that a full correction is not necessary to break the illusion.

1.3.1 Study Design

The study aims to determine the range of "optimal" transformation weights for each transformation type. Psychophysics methodology typically approaches threshold estimation by using the method of adjustment (**?**), where stimuli are provided showing states both above and below the hypothesized optimal value and participants adjust the stimuli until the stated goal is met (in this case, until the lines appear to have equal length). It is expected that there will be a difference in user-reported values from below and from above, and these values are typically averaged to produce a single threshold value. Beyond averaging these values, we use a mixed

Figure 1.15: Overview of possible starting weights. Weight values are discrete, but staggered so as to provide fine-grained adjustments around 0.6 and more coarse discriminatory information toward the outside.

model to compare user responses for different starting points in a more continuous fashion, incorporating some of the advantages of the method of constant stimuli to more robustly estimate the range of optimal transformation weights. For a review of general psychophysics methodology, the method of adjustment, and the method of constant stimuli, see **?**.

The study is set up as a fractional factorial design of correction type (x or y correction) and starting weight w_0. Each participant is asked to evaluate a total of twelve situations, six of each correction type. Starting weights were chosen as follows: each user was given a trial of each type starting at 0 and 1. The remaining four trials of each type had starting weights chosen with equal probability from 0.25 to 0.75 (see figure **??**). We decided to have a higher coverage density for starting weights around 0.6 after a pilot study indicated a preference for that value. Using a distribution with a wide coverage allows us to more fully explore the space of plausible weights w while focusing on the $(0, 1)$ interval and enabling precise estimation of the optimal weight in the region indicated by the pilot study.

A trial begins with the presentation of a graph at the chosen starting weight w_0. Participants adjust the graph using increment and decrement buttons. A trial ends with the participant clicking the 'submit' button, at which point the weight for the final adjustment is recorded. This provides a clear starting value and ending value, allowing us to assess the range of optimal values for each participant. In addition to starting weight, correction type, and anonymized user-specific data (partial IP address, hashed IP address, and hashed browser characteristics), each incremental user chosen weight is recorded with a corresponding timestamp. The user-specific browser data is sufficient to provide a 'fingerprint' to distinguish and recognize individual users (or rather their computer settings) in an anonymous fashion.

Each participant is provided with two initial "training" trials in which the graph of the underlying mean function is superimposed on the line segments to give participants some idea of the function the lines represent. This approach was taken to reduce incidences of extremely high correction values under the X transformation, as large adjustment values do not change

Figure 1.16: Transformation weights outside of the intervals $[-2.5, 3.5]$ for y and $[-2, 2]$ for x produce figures which do not maintain the underlying function shape (in x) or which are composed of extremely uneven length lines (in y). Trials with final results that were more extreme than these examples were excluded from the analysis.

the impression of same line length, but the resulting function bears little resemblance to a sine function, see figure **??** for examples of overcorrection.

1.3.2 Results

Participants were recruited from Amazon Mechanical Turk and the reddit community.

As this study was conducted outside a laboratory setting, we can not gauge a participant's willingness to follow the guidelines and put in their best effort. This, besides potential technical issues (server outage, speed of response) make a careful selection of data going into the analysis unavoidable. The following exclusion criteria were used:

- Participants did not interact with the applet: we required participants to use the adjustment at least once in order to include data for this trial (592 trials removed).

- Participants finished fewer than four trials: while participants were asked to complete twelve trials, some did not finish all of those. In order to stabilize predictions of random effects, participants' data were excluded if there were fewer than four trials (78 out of a total of 203 participants).

- Out-of-bounds results: weights leading to severely over- or undercorrected results were excluded from the analysis. For trials to adjust Y-values, weights outside of $[-2.5,\ 3.5]$ show dramatically unequal line lengths; weights from X-transformations outside the range of $[-2,\ 2]$ do not preserve the underlying function shape and concavity. Figure **??** shows results at the threshold of acceptability. Only more severely distorted results were excluded from the analysis (12 of the X and 5 of the Y trials out of 1227 trials remaining after application of other criteria).

The following analysis is based on the cleaned data, consisting of 125 participants with 1210 valid trial results. The psychophysics model shown in figure **??** is based on weighted averages

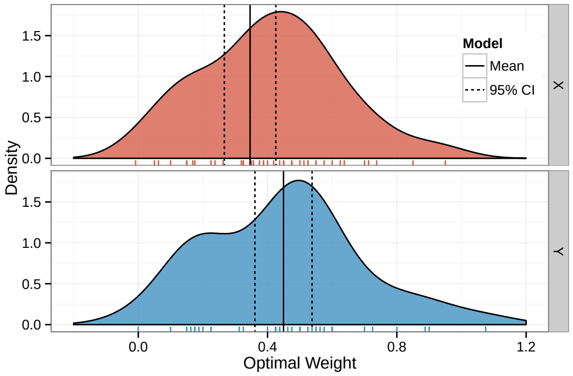


Figure 1.17: Estimated density of participant-level means using the standard psychophysics method of limits analysis. The overall means are both near 0.4, however, there is quite a bit of user-level variability.

(by adjustment type) of all trials with starting weights $w_0 = 0$ and 1.

According to this analysis, the optimum transformation value for x is 0.35, and the optimum transformation value for y is 0.45. Figure **??** shows the estimates and 95% Wald intervals for the mean, as well as estimated density of participant-level responses.

While these results suggest that the transformation is useful and that complete transformation is not necessary, we can get more precise bounds on the range of acceptable transformation weights using a linear model that can incorporate starting points other than 0 and 1, and at the same time allow for user-specific variability.

In order to account for user-level variability, we fit a random effects model for the adjusted weight value as a function of starting weight and trial type, with a random intercept for each participant.

Let W_{ij} denote the final adjustment to weight by participant i, $1 \leq i \leq 125$, on trial j, $1 \leq j \leq n_i$. We model the final weight W_{ij} as a function of the correction type $T(i,j)$ (where $T(i,j) \in \{X, Y\}$), and starting weight X_{ij}, with a random intercept for participant to account for subject-specific ability:

$$W_{ij} = \alpha_{T(i,j)} + \beta X_{ij} + \gamma_{i,T(i,j)} + \epsilon_{ij} \tag{1.11}$$

$$\gamma_{iX} \overset{\text{i.i.d.}}{\sim} N(0, \eta_X^2), \ \gamma_{iY} \overset{\text{i.i.d.}}{\sim} N(0, \eta_Y^2),$$

$$\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \text{ and } \mathrm{Cov}(\gamma, \epsilon) = 0$$

$\alpha_{T(i,j)}$ is either α_X or α_Y, describing the lower threshold of the acceptable range for each of the types of correction, while $\alpha_X + \beta$ and $\alpha_Y + \beta$ describe the upper thresholds for the respective correction.

We can therefore interpret β as the length of the interval of plausible weights. Additionally, this allows the interpretation of the quantity $(\alpha_* + \beta/2)$ as equivalent to the estimate of the optimal weight based on the psychophysics methodology.

The fitted model parameters are shown in tables **??** and **??**.

Transformation	Threshold	Parameter	Estimate	95% C.I.
X	Lower	α_X	0.097	(0.045, 0.150)
	Upper	$\alpha_X + \beta$	0.625	(0.570, 0.682)
Y	Lower	α_Y	0.143	(0.097, 0.188)
	Upper	$\alpha_Y + \beta$	0.671	(0.626, 0.718)

Table 1.1: Fixed effect estimates of model (**??**) for the boundaries for reasonable weights. In parentheses, 95% parametric bootstrap confidence intervals are given based on model (**??**) (N=1000).

Groups	Correction	Parameter	Estimate	95% C.I.
Participant	X	η_X	0.171	(0.167, 0.247)
Participant	Y	η_Y	0.145	(0.107, 0.179)
Residual		σ	0.304	(0.290, 0.317)

Table 1.2: Overview of random effects for model (**??**), including 95% confidence intervals based on parameteric boostrap results (N=1000).

Table **??** gives an overview of the variance estimates. 95% confidence intervals are, based on 1000-fold parametric bootstrap of model **??**. All variance components are significant and relevant; variability within a single individual's trials is about half the size of variability across participants.

We use parametric bootstrap to generate responses for each correction type and each participant from the model, which we use to both create user-level densities, population-level densities, and bootstrap intervals for model parameters.

The variability of the random effects for each trial type is similar; but the model benefits significantly from allowing separate random effects for individual's variability by correction type (0.1452394 and 0.1705474 for Y and X transformations, respectively, as opposed to 0.3044344 for the overall variability). The interaction between starting weight and trial type was not significant, however, and was thus removed from the model (p-value = 0.9009749).

Figure 1.18: Simulation results from the fitted model, facetted by correction type. Fixed effects results are shown as histograms; the red values display the results when starting from an uncorrected plot and are concentrated around $w = 0.1$ for X and $w = 0.14$ for Y; the blue values represent user-chosen weights when starting from a fully corrected plot and are concentrated around $w = 0.63$ for X and $w = 0.67$ for Y. Additionally, 95% bootstrap intervals are shown as horizontal line segments above the histograms; these intervals are for the lower and upper bounds of the "preferred weight interval" tested in the experiment. User-level density curves show the individual variability around fixed effects α_* and $\alpha_* + \beta$.

Figure **??** gives an overview of the relationship between starting weights and user-preferred weight values. Higher starting weights are associated with higher user-submitted values, and lower starting weights are associated with lower user-submitted values.

The ranges of optimal weights are similar under both transformations. Boundaries for the X transformation are slightly lower than boundaries for Y.

Bootstrap simulations for each of the coefficients suggest that the range of optimal w is between 0.098 and 0.625 for x and 0.142 and 0.67 for y, where the lower value is the estimate starting at $w = 0$ and moving up, and the upper value is the estimate starting at $w = 1$ and moving down. This suggests that either correction is preferrable to an uncorrected graph, and that a weighted correction is preferrable to the fully corrected graph, as neither 0 nor 1 is contained in any overall interval. In addition to showing the strength of the correction, this experiment also demonstrates the strength of the illusion itself: a correction appears more uniform than the uncorrected values, even though the corrected values are not uniform and the uncorrected values are completely uniform.

1.4 Application: US Gas Prices

Figure **??** shows daily gas prices for a time frame between 1995 to 2014 as published in the Energy Information Administration's historical database of gas prices (**?**). This data includes prices for all three grades of gasoline as well as two chemical formulations which are sold in different geographic areas across the United States (for more information, see (**?**)).

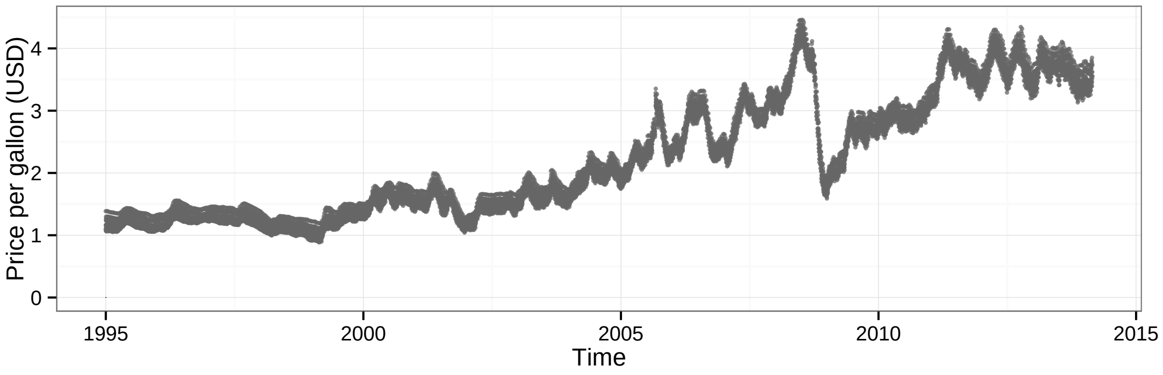


Figure 1.19: US Gas prices from 1995 to 2014. Gas prices steadily increase over the time frame, with some dramatic short-term developments. Peaks and troughs seem to exhibit more variability in daily prices than times of dramatic changes. This is an effect of the sine-illusion, which hides a fairly steady increase in variance in daily gas prices over time.

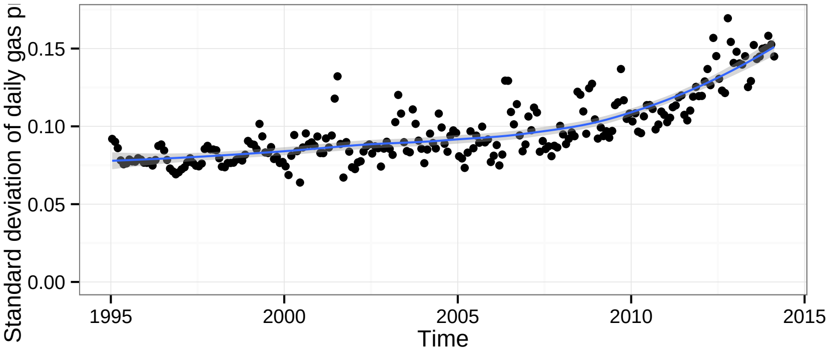


Figure 1.20: Standard deviation of daily gas prices between 1995 and 2014. The doubling of the standard deviation over the time frame is masked in figure ??.

There is a clear increase in daily gas prices over time as well as several dramatic price changes. These developments mask the steady increase in variance shown in figure ??. Instead, we perceive an increase in variability in the frequent ups and downs along the overall trend. In particular, the strong decrease in gas prices at the end of 2008 seems to be associated with a low variance. This is an effect of the sine illusion, and the actual variability in Oct 2008 is higher than previous months. In order to better judge variability along the trendline we applied the two different corrections to this data.

For either of the corrections we use a trendline fit based on smoothing splines, which provides the necessary first and second derivatives.

Figure ?? shows the results from the X transformation applied to the gas prices. The figure on top is a fully corrected version, while the one below only uses $w = 0.36$, the midpoint of the

Figure 1.21: Gas price data corrected using the X transformation with $w = 1$ and with $w = 0.36$.

range of experimentally determined acceptable values, for the transformation. At $w = 1$, the transformation is severe, but it becomes clear that the variance between 1995 and 2000 is lower than it is between 2009 and 2014. When $w = 0.36$, the transformation is much less noticable but yields a near-constant absolute slope of the fitted line.

The minor effect of the weighted transformation on individual x-values contrasts with the effectiveness of the transformation in reducing the illusion; this is best seen in the fitted line, which is distinctly (piecewise) curved in the uncorrected data and appears to be much more piecewise linear in the corrected data, even at the reduced weighted value.

Similar to the X transformation, the Y transformation highlights local fluctuation in the variability of daily gas prices much more than the untransformed data. Figure **??** shows Y transformations for the data. Again, we show a full transformation (top) and a transformation based on the midpoint of the previously determined acceptable region of $w = 0.40$. in the full transformation it is clear that the variance is nearly constant between 1995 and 2000 and then begins to increase with the price of gas. When $w = 0.40$, the transformation is much less noticable, and the resulting y-axis scale is much more similar to the uncorrected data.

Figure 1.22: Gas price data corrected using the Y transformation with $w = 1$ and with $w = 0.40$.

1.5 Conclusions

The sine illusion is a persistent and powerful illusion that is very difficult to counteract without modifying the visual stimulus directly. While systematically modifying the data is uncommon in the statistical world, this approach is not out of place in the visual arts or architecture; as far back as 400 BC the builders of the Parthenon ensured a straight appearance of the columns from afar, by widening columns at the center, thereby counteracting the effects of the Hering illusion (**??**). Similarly, painters often exaggerate color hues used in shadows to account for color constancy in the brain. The systematic modifications we suggest here are also comparable to chloropleth maps, which scale a region's area based on some other variable.

We cannot counteract the illusion and represent the data visually without an intervention that is drastic enough to counteract the three-dimensional context the sine-illusion induces. The proposals in this paper for transformations in x and y provide the means to temporarily correct the data as a diagnostic measure, perhaps using an applet or R package for that purpose. These corrections are significant not only because of their implications for statistical graphics, but because previous attempts to resolve optical illusions using geometry have not met with success (**?**). These corrections are only a first step and could be improved upon; currently, the corrections break down for extreme (secant) values, but multiple iterations of the correction procedure will likely resolve some of these issues (though iteration removes the convenience of a functional form for the transformation). Similarly, the y corrections proposed here extend the line lengths (or for actual data, increase the deviation from the smooth line) – some normalization might make the necessary corrections less noticeable.

Our primary goal is to raise awareness of the illusion and its implications for statistics; the use of plots to guide the modeling process can leave us vulnerable to overlooking changes in the variance due to the illusion. While best practice has been to plot the residuals separately, this removes the context of the data and is not practical before there is a model. In addition, viewer attention spans may be limited if multiple graphs are presented. The proposed transformations require only a nonparametric smooth, maintain the context of the data, and are readily interpretable.

The data for this study was collected with approval from IRB-ID 13-257.