# DATA 227

## Uncertainty in Data Visualizations, Part II

2022-11-10

# Vocabulary Review



- Population: an entire group of people, objects, observations, etc. that satisfy some set of characteristics.
  - e.g., all adults living in the United States, all cans of baby formula produced at a particular plant.
- Sample: a subset of people, objects, observations, etc. coming from a particular population.

# Vocabulary Review



- Parameter: some characteristic (usually numeric) of a population.
  - e.g., the average height of all adults living in the United States, the proportion of cans of baby formula tainted by *Chronobacter*.
- Statistic: a value calculated numerically from a sample. Typically, this value is intended to estimate a parameter.

# Statistics



Statistic: a value calculated numerically from a sample. Typically, this value is intended to estimate a parameter.

- Sample mean or average, $\bar{x}$, intended to estimate a population mean $\mu$
- Sample standard deviation, $s$, intended to estimate a population standard deviation, $\sigma$
- Sample proportion $\hat{p}$, intended to estimate a population proportion $p$.
- **Sample correlation, $r$, intended to estimate a population correlation coefficient $\rho$ (see also, regression coefficients)**

$\bar{x}$, $s$, $r$, and $\hat{p}$ are all known as point estimates–they estimate the parameters with a single value.

# Visualizing Uncertainty



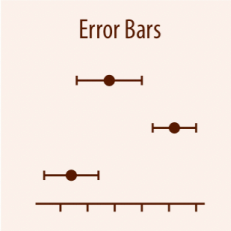
"When we see a data point drawn in a specific location, we tend to interpret it as a precise representation of the true data value. It is difficult to conceive that a data point could actually lie somewhere it hasn't been drawn. Yet this scenario is ubiquitous in data visualization. Nearly every data set we work with has some uncertainty, and whether and how we choose to represent this uncertainty can make a major difference in how accurately our audience perceives the meaning of the data."

Claus Wilke, Fundamentals of Data Visualization, Visualizing Uncertainty

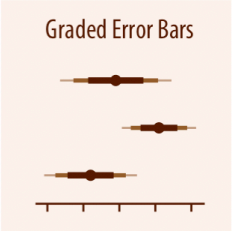
# Visualizing Uncertainty in Means and Proportions



- **Error bars** are meant to indicate the range of likely values for some estimate or measurement (a.k.a., confidence intervals). They extend horizontally and/or vertically from some reference point representing the estimate or measurement. Reference points can be shown in various ways, such as by dots or by bars.
- **Graded error bars** show multiple ranges at the same time, where each range corresponds to a different degree of confidence. They are in effect multiple error bars with different line thicknesses plotted on top of each other.





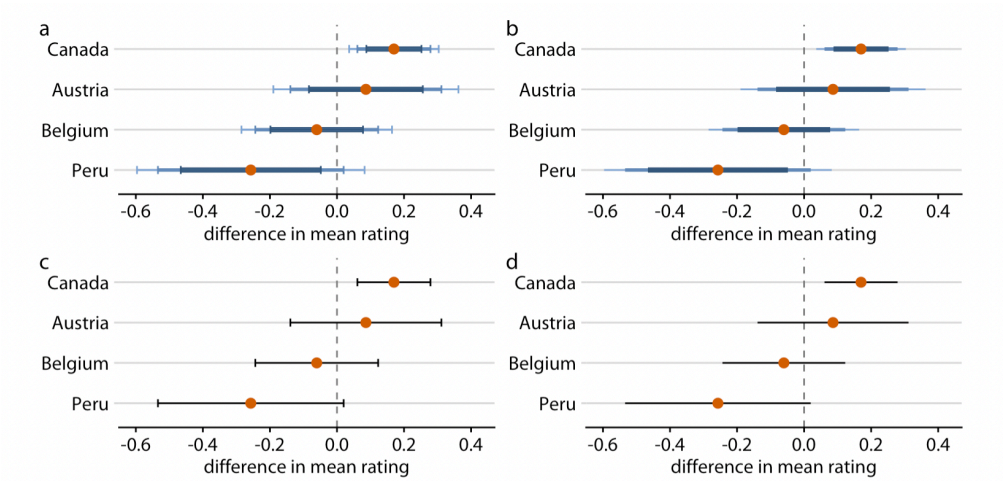
# More on Error Bars 1



- More variations are possible!
  - For example, we can draw error bars with or without a cap at the end.
- There are advantages and disadvantages to all these choices.
  - Graded error bars highlight the existence of different ranges corresponding to different confidence levels.
  - However, the flip side of this additional information is added visual noise.
  - Whether to draw error bars with or without cap is primarily a question of personal taste. A cap highlights where exactly an error bar ends, whereas an error bar without cap puts equal emphasis on the entire range of the interval.
  - Also, again, caps add visual noise, so in a figure with many error bars omitting caps may be preferable.

# More on Error Bars 2





# Visualizing Uncertainty in Means and Proportions 2

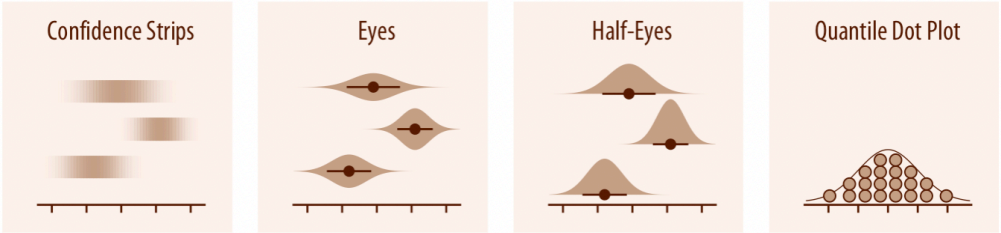


To achieve a more detailed visualization than is possible with error bars or graded error bars, we can visualize the actual confidence or posterior distributions (Bayesians only).

- **Confidence strips** provide a clear visual sense of uncertainty but are difficult to read accurately.
- **Eyes** and **half-eyes** combine error bars with approaches to visualize distributions, and thus show both precise ranges for some confidence levels and the overall uncertainty distribution.
- A **quantile dot plot** can serve as an alternative visualization of an uncertainty distribution and can be easier to read than a violin or ridgeline plot.

# Visualizing Uncertainty in Means and Proportions 3





# Confidence Strips, Distributions, and Dotplots 1



- As an alternative to error bars we could draw confidence strips that gradually fade into nothing.
- Confidence strips better convey how probable different values are, but they are difficult to read.
- We would have to visually integrate the different shadings of color to determine where a specific confidence level ends.
- It is difficult to visually integrate the area under the curve and to determine where exactly a given confidence level is reached. This issue can be somewhat alleviated, however, by drawing quantile dotplots.

# Confidence Strips, Distributions, and Dotplots 2



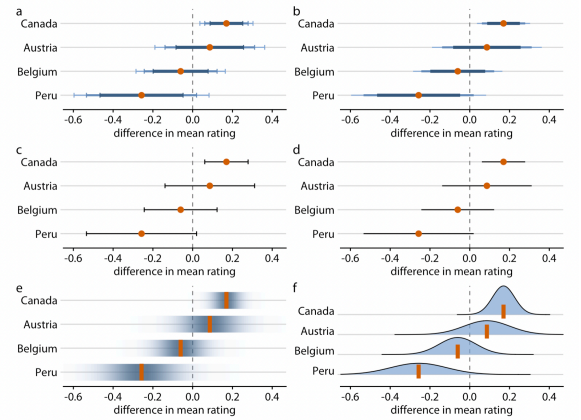



Figure 16.9: Mean chocolate flavor ratings for manufacturers from four different countries, relative to the mean rating of U.S. chocolate bars. Each panel uses a different approach to visualizing the same uncertainty information. (a) Graded error bars with cap. (b) Graded error bars without cap. (c) Single-interval error bars with cap. (d) Single-interval error bars without cap. (e) Confidence strips. (f) Confidence distributions.

# Visualizing Uncertainty



"When we see a data point drawn in a specific location, we tend to interpret it as a precise representation of the true data value. It is difficult to conceive that a data point could actually lie somewhere it hasn't been drawn. Yet this scenario is ubiquitous in data visualization. Nearly every data set we work with has some uncertainty, and whether and how we choose to represent this uncertainty can make a major difference in how accurately our audience perceives the meaning of the data."

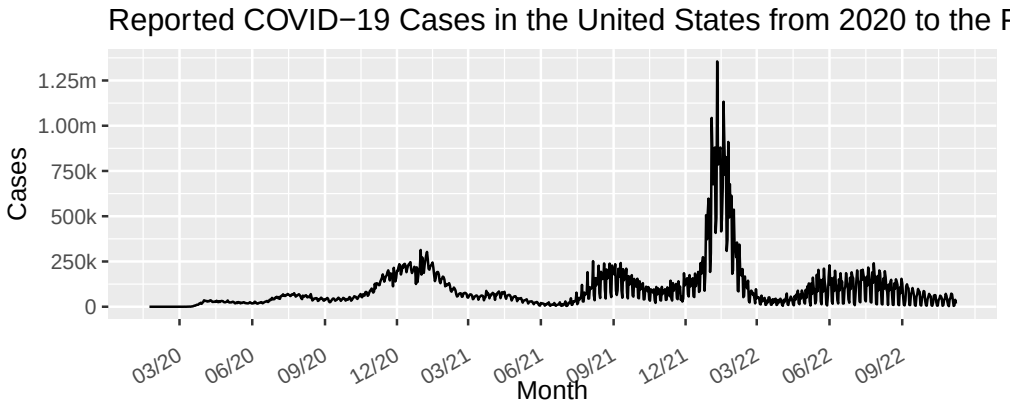
Claus Wilke, Fundamentals of Data Visualization, Visualizing Uncertainty

- This is especially difficult when plotting trend lines!

# Visualizing Trends: Line Plots 1



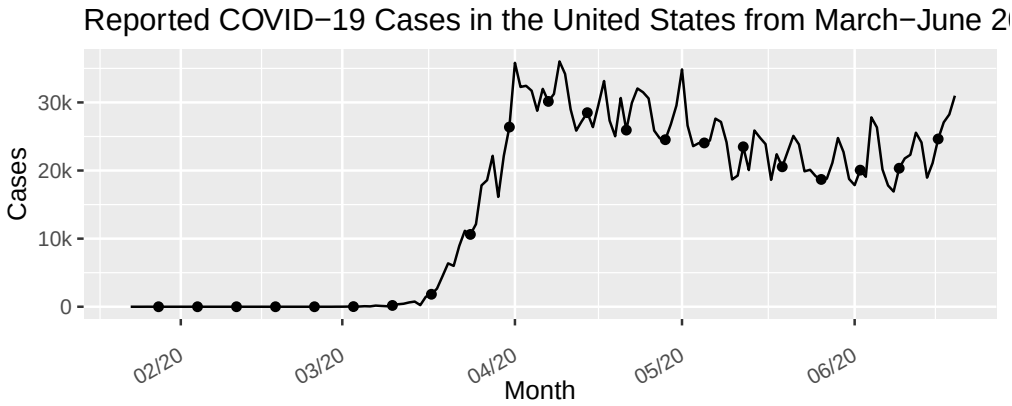
- The simplest way to visualize a trend is a line plot.



# Visualizing Trends: Line Plots 2



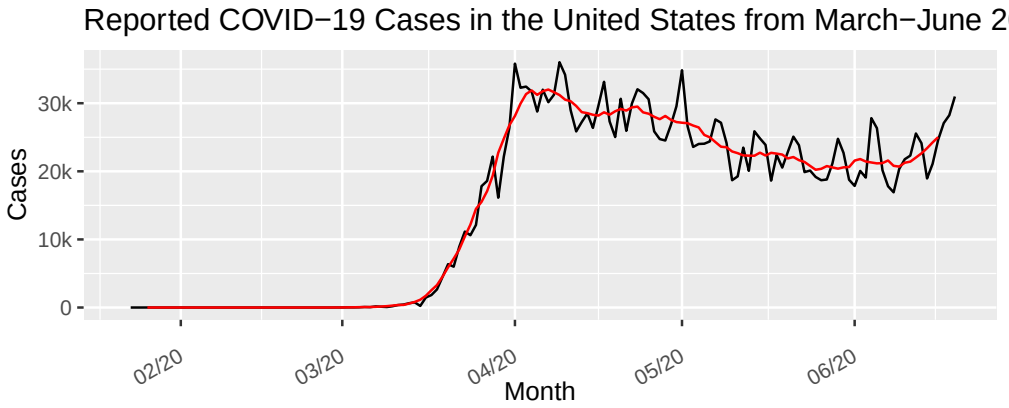
- Line points can be very jagged, and you might be "overfitting" the trend.



# Smoothing: Moving Averages 1



- How can we visualize these longer-term trends while de-emphasizing the less important short-term fluctuations? Moving averages.



# Smoothing: Moving Averages 2



- The moving average is the most simplistic approach to smoothing. However, it has some obvious limitations.
  - It results in a smoothed curve that is shorter than the original curve. Parts are missing at either the beginning or the end or both. And the more the time series is smoothed (i.e., the larger the averaging window), the shorter the smoothed curve.
  - Even with a large averaging window, a moving average is not necessarily that smooth. It may exhibit small bumps and wiggles even though larger-scale smoothing has been achieved.

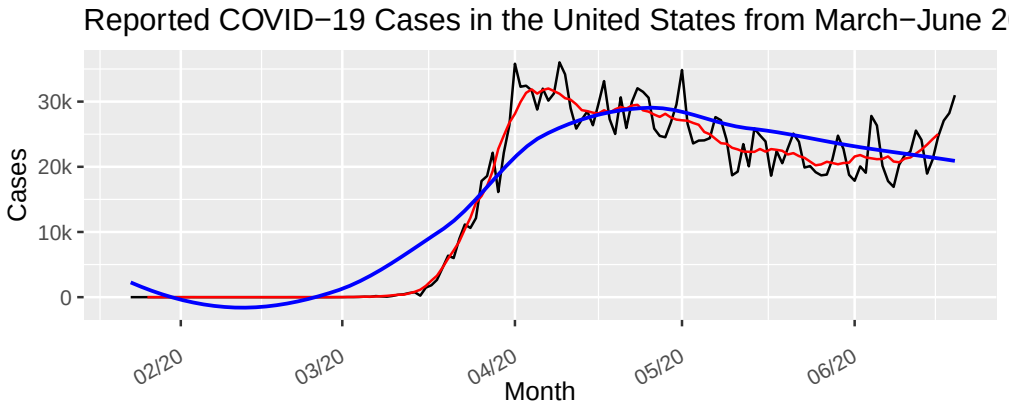
# Smoothing: LOESS 1



- Statisticians have developed numerous approaches to smoothing that alleviate the downsides of moving averages.
- These approaches are much more complex and computationally costly, but they are readily available in modern statistical computing environments.
- One widely used method is LOESS (locally estimated scatterplot smoothing, W. S. Cleveland (1979)), which fits low-degree polynomials to subsets of the data.
- Importantly, the points in the center of each subset are weighted more heavily than points at the boundaries, and this weighting scheme yields a much smoother result than we get from a weighted average.

# Smoothing: LOESS 2

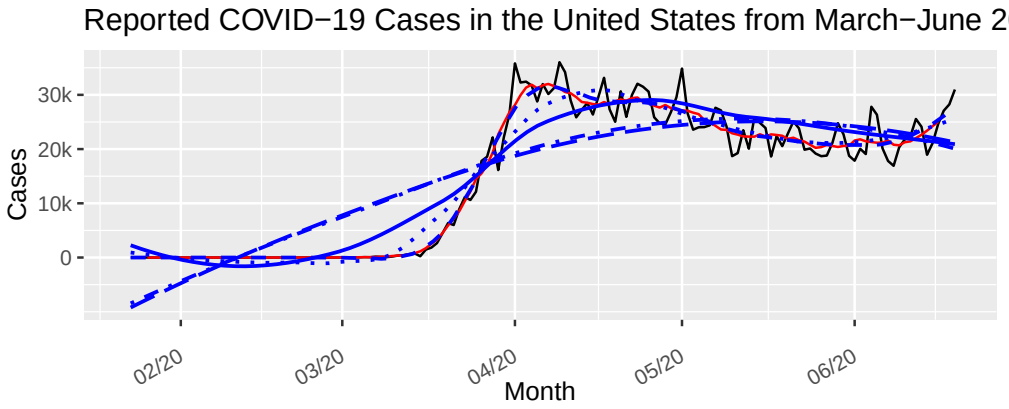




# Smoothing: LOESS 3



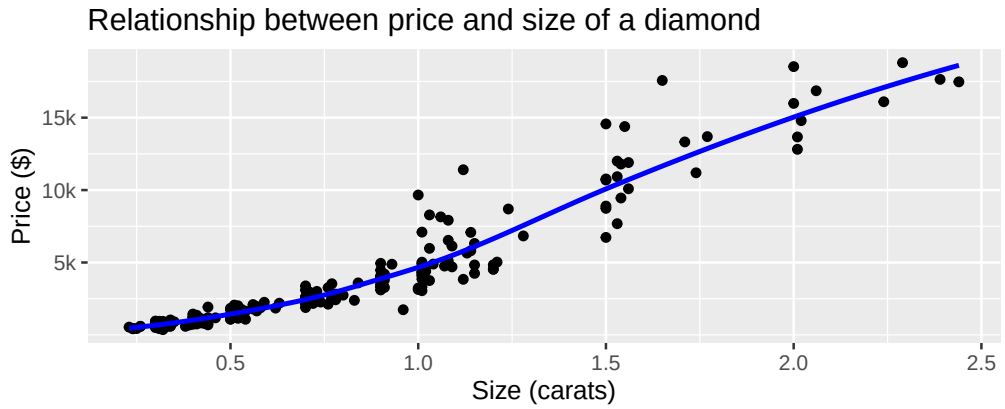
- The smoothness of a LOESS curve can be tuned by adjusting a parameter, and different parameter choices would have produced different LOESS curves.



# Smoothing: LOESS 4



Importantly, LOESS is not limited to time series. It can be applied to arbitrary scatter plots.



# Smoothing: LOESS 5



- LOESS is a very popular smoothing approach because it tends to produce results that look right to the human eye.
- However, it requires the fitting of many separate regression models. This makes it slow for large datasets, even on modern computing equipment.

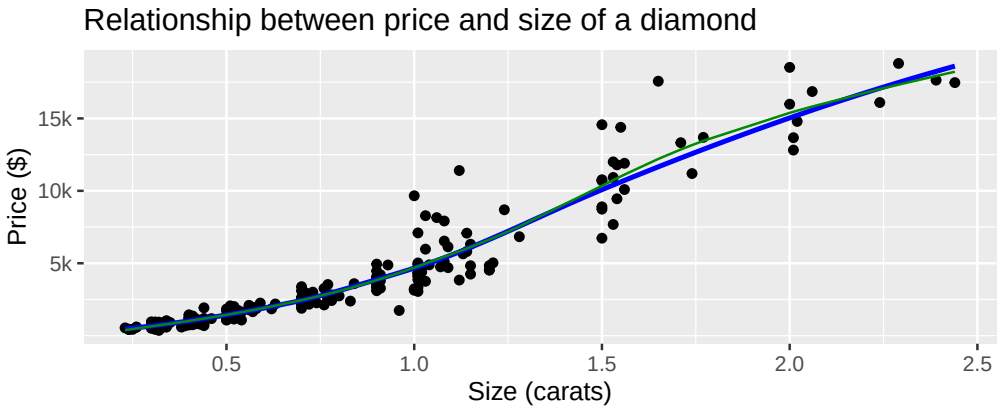
# Smoothing: Splines 1



- As a faster alternative to LOESS, we can use spline models.
- A spline is a piecewise polynomial function that is highly flexible yet always looks smooth.
  - When working with splines, we will encounter the term knot. The knots in a spline are the endpoints of the individual spline segments. If we fit a spline with $k$ segments, we need to specify $k + 1$ knots. While spline fitting is computationally efficient, in particular if the number of knots is not too large, splines have their own downsides.
  - Most importantly, there is a bewildering array of different types of splines, including cubic splines, B-splines, thin-plate splines, Gaussian process splines, and many others, and which one to pick may not be obvious.
- The specific choice of the type of spline and number of knots used can result in widely different smoothing functions for the same data.

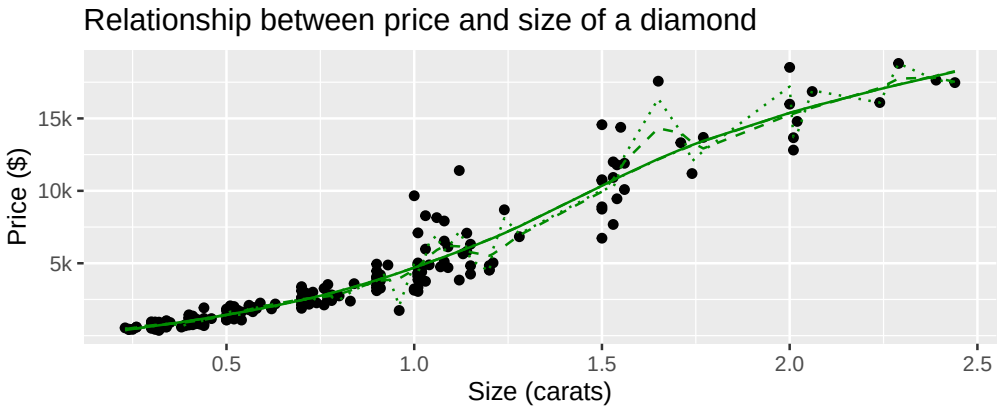
# Smoothing: Splines 2





# Smoothing: Splines 3





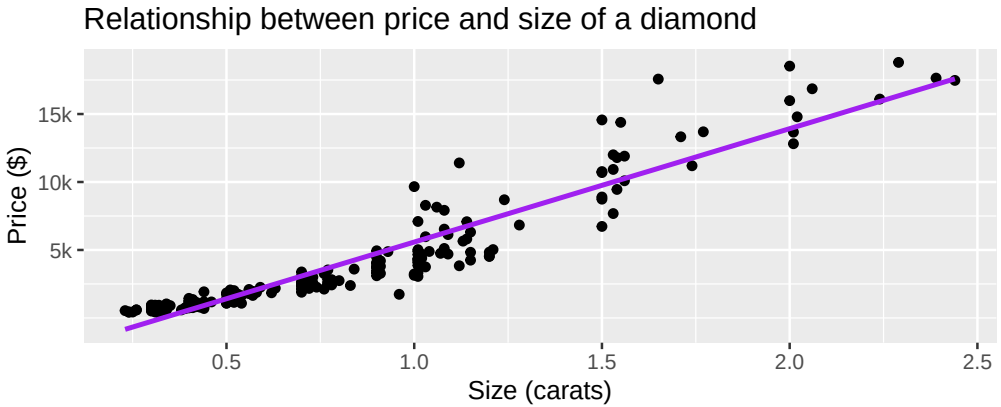
# Smoothing: Defined Functional Forms 1



- The behavior of general-purpose smoothers can be somewhat unpredictable for any given dataset.
- These smoothers also do not provide parameter estimates that have a meaningful interpretation.
- Therefore, whenever possible, it is preferable to fit a curve with a specific functional form that is appropriate for the data and that uses parameters with clear meaning.
  - Linear
  - Polynomials
  - Exponential

# Smoothing: Defined Functional Forms 2






- The slope can be interpreted here as, "for every one carat increase in size, the price is expected to increase by $8,338.

# Visualizing Uncertainty in Trends 1

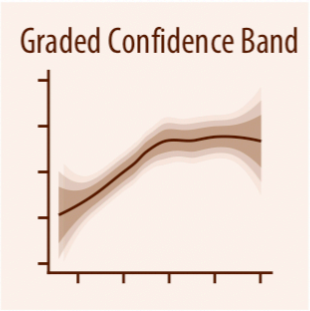


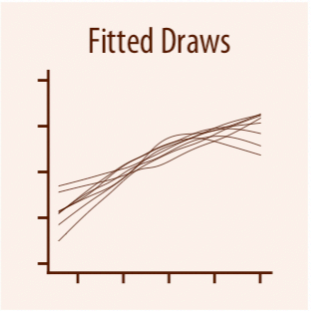
- We've talked about using scatterplots and correlation to talk about relationships–but correlation (and the coefficients of smoothers) are also statistics, and therefore have associated uncertainty.
- For smooth line graphs, the equivalent of an error bar is a confidence band. It shows a range of values the line might pass through at a given confidence level.

# Visualizing Uncertainty in Trends 2



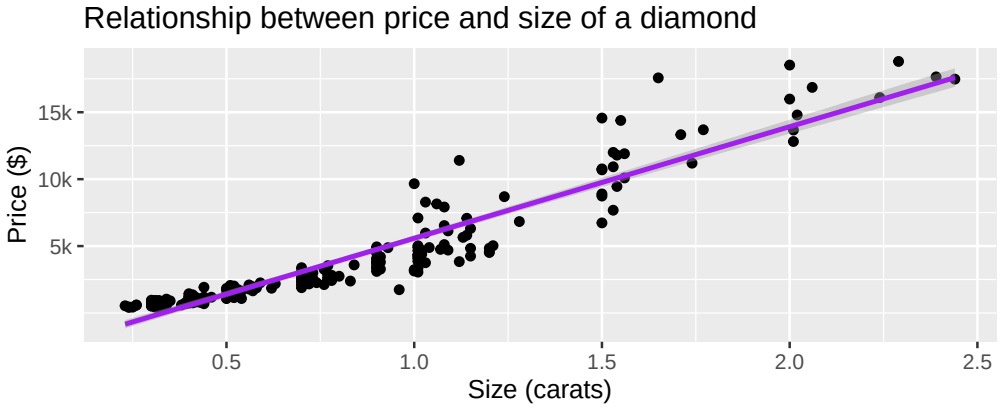
- As in the case of error bars, we can draw graded confidence bands that show multiple confidence levels at once. We can also show individual fitted draws in lieu of or in addition to the confidence bands.

# Smoothing and Uncertainty 1

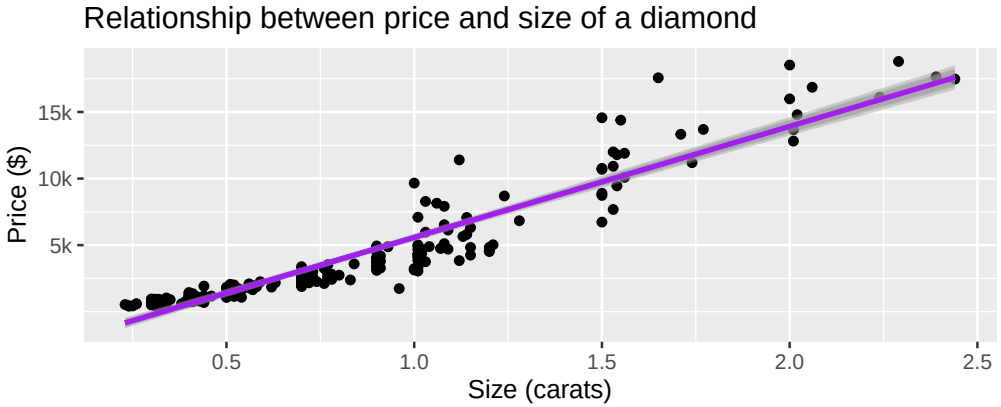




# Smoothing and Uncertainty 2







# Sources

Wilke, Claus O. Fundamentals of data visualization: a primer on making informative and compelling figures. O'Reilly Media, 2019.