DATA 227

Uncertainty in Data Visualizations

2022-11-08

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.

Distributions 1

- If you have access to the entire population, there is no uncertainty!
- Instead, you are more likely interested in the distribution of the data.
- A distribution can be loosely defined as a summary of the values a variable can take and how often they occur.
  - Bar charts, histograms, and boxplots are all ways of visualizing distributions.

Distributions 2

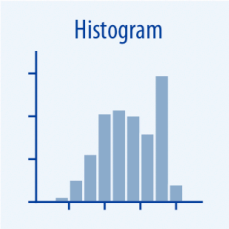
- With regards to distributions of numeric variables, there are a few characteristics of distributions we are typically interested in.
  - Location (a.k.a., the center of a distribution)
  - Scale (a.k.a., the spread or variability in a distribution)
  - *Shape*
  - *Outliers/unusual values*

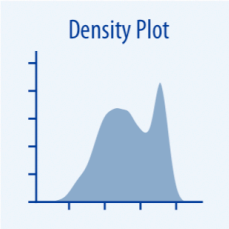
Variability

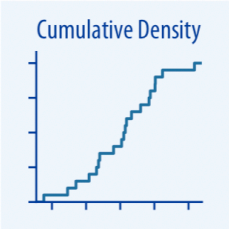
- **Variability is natural to any distribution for a random variable. It refers to the spread or dispersion inherent to a population.**

Visualizing Distributions 1

- **Histograms** and **density plots** provide the most intuitive visualizations of a distribution, but require arbitrary parameter choices and can be misleading.
- **Cumulative densities** and **quantile-quantile (q-q) plots** always represent the data faithfully, but can be more difficult to interpret.

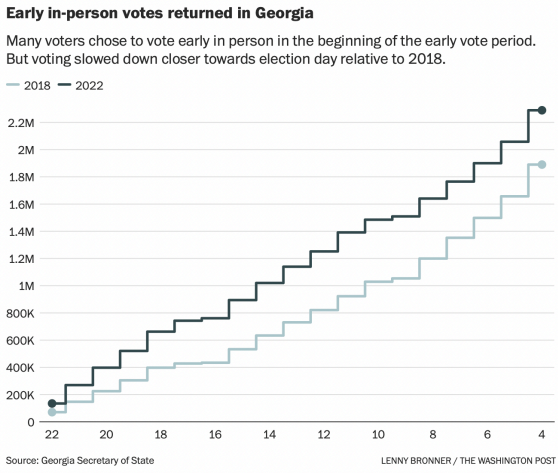






Fundamentals of Data Visualization

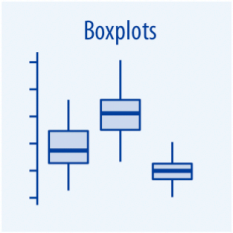
Cumulative Density

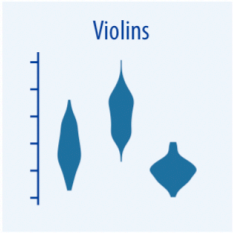


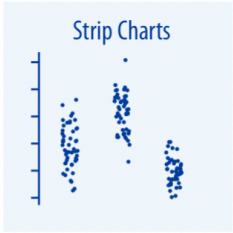
Number of early votes cast surpasses early-vote total in 2018 midterm election, WaPo

Visualizing Distributions 2

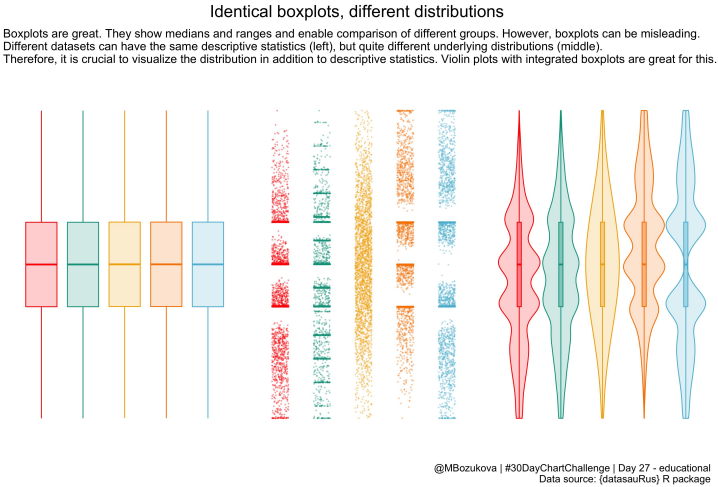
- **Boxplots**, **violins**, **strip charts**, and **sina plots** are useful when we want to visualize many distributions at once and/or if we are primarily interested in overall shifts among the distributions.



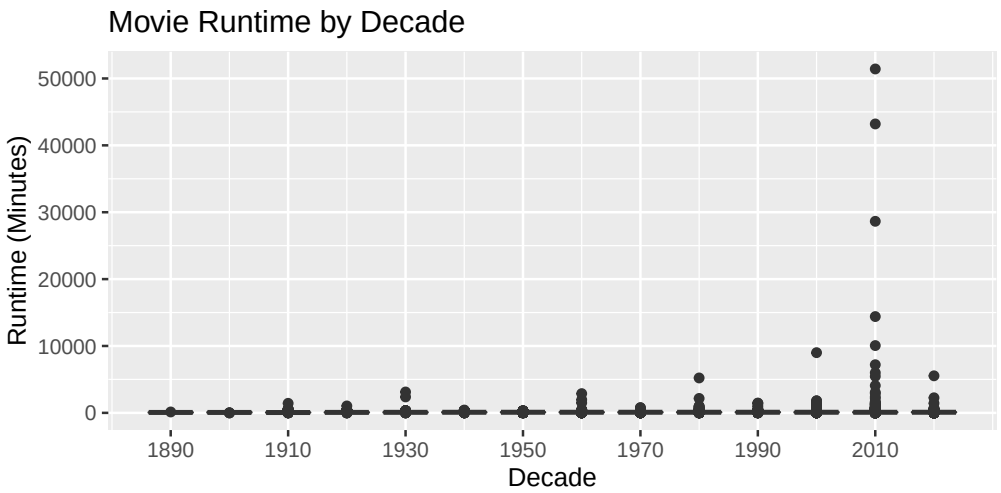




Boxplots, Violin Plots, Strip Charts

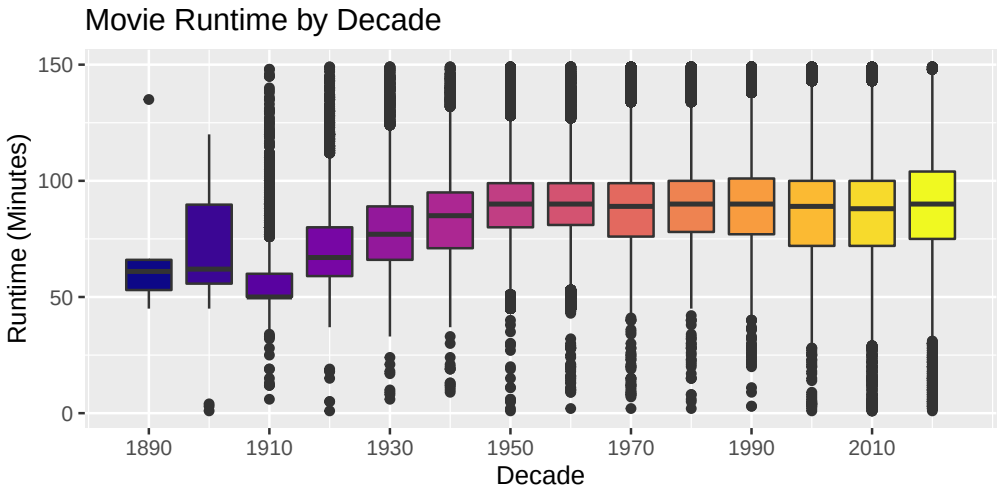


Box Plots 1

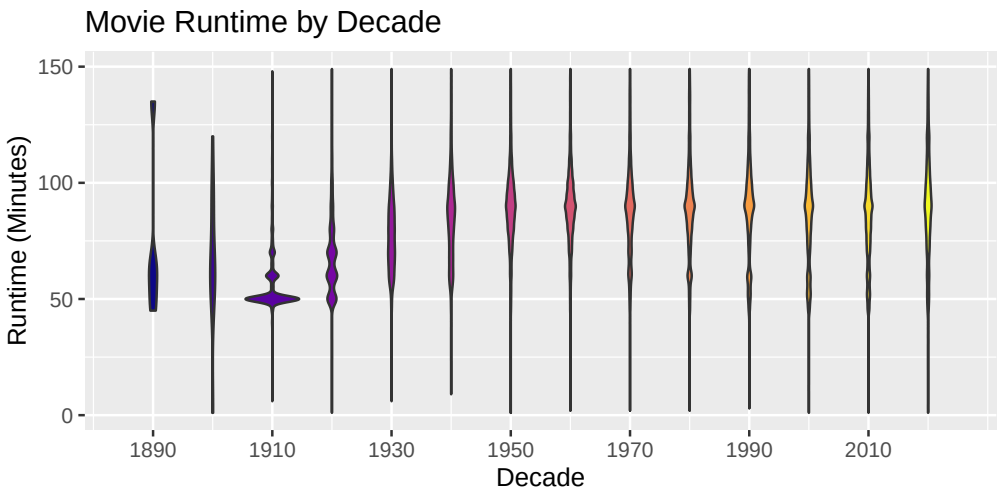


IMDB Datasets

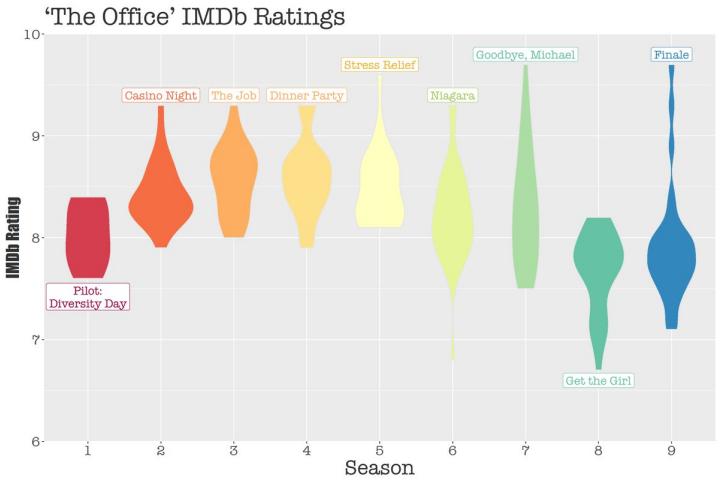
Boxplots 2


Movie Runtime by Decade
Runtime (Minutes)
150
100
50
0
1890
1910
1930
1950
1970
1990
2010
Decade

Violin Plots



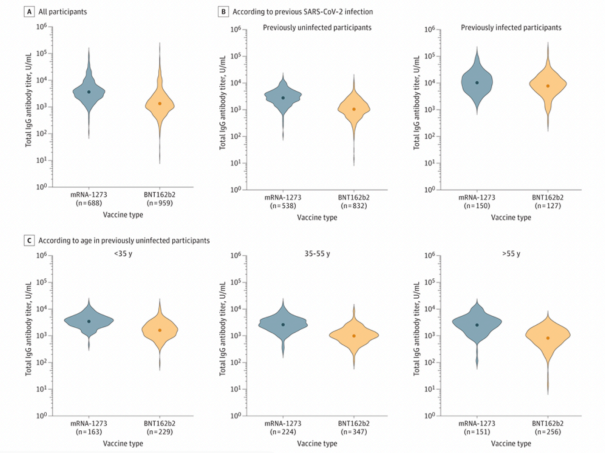
Violin Plots: In the Wild 1



@_isabellamb

Violin Plots: In the Wild 2

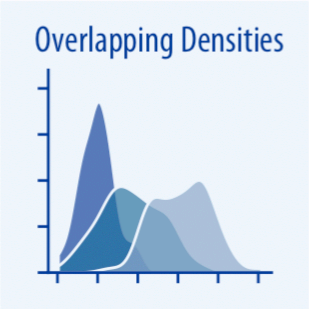
Figure. Humoral Immune Response Following SARS-CoV-2 mRNA Vaccination

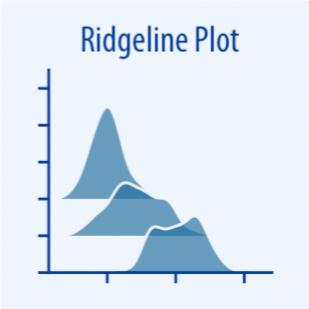


Comparison of SARS-CoV-2 Antibody Response Following Vaccination With BNT162b2 and

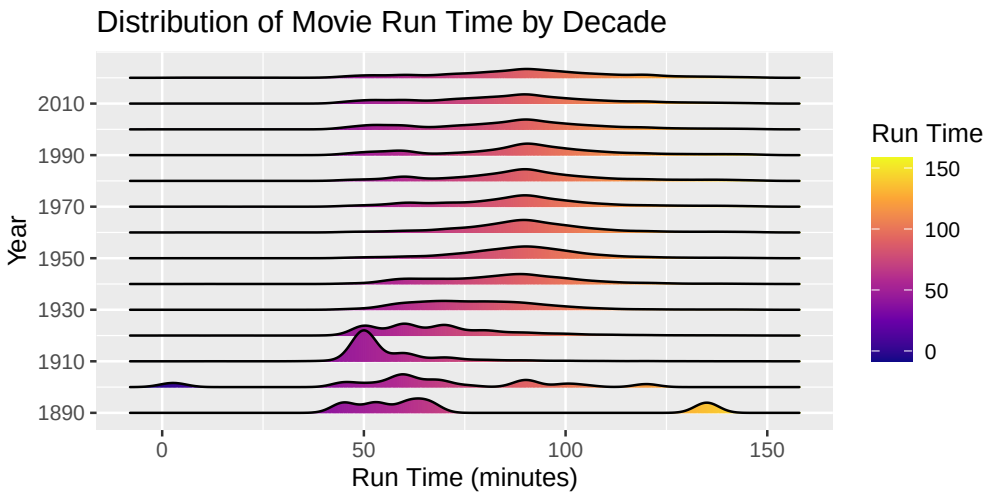
Visualizing Distributions 3

- **Stacked histograms** and **overlapping densities** allow a more in-depth comparison of a smaller number of distributions, though stacked histograms can be difficult to interpret.
- **Ridgeline plots** are often useful when visualizing very large numbers of distributions or changes in distributions over time.

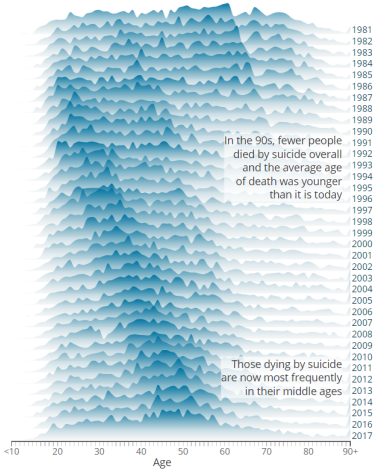




Ridgeline Plots



Ridge Plots: In the Wild 1

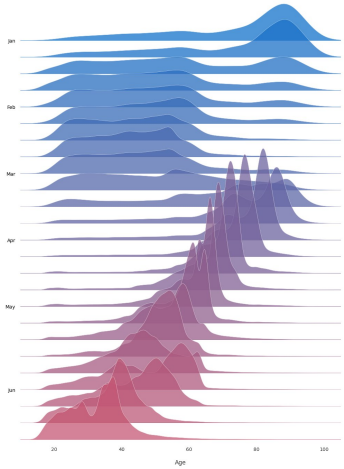


Source: Office for National Statistics

Ridge Plots: In the Wild 2

Kenneth KHW

Ridge Plots: In the Wild 3



MathiasLeroy_

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 μ
- Sample standard deviation, s, intended to estimate a population standard deviation, σ
- Sample correlation, r, intended to estimate a population correlation coefficient ρ (see also regression coefficients)
- Sample proportion $\hat{p}$, intended to estimate a population proportion p.

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

Defining Uncertainty

- Because of variability in a population, samples from that population are incredibly unlikely to be identical.
- Because samples are unlikely to be identical, your statistic, calculated from the sample, will vary from sample to sample.
- We can quantify this variability–known as sampling variability, sampling error, or uncertainty.
- Specifically, we use the standard error, which tells us how precisely we have determined a parameter estimate, to build interval estimators.

Interval Estimators

- To address uncertainty, we often like to use interval estimators rather than point estimates.
- Interval estimators are ranges of values, and are built by using statistics and their sampling distributions.
  - Frequentist Inference: Confidence Intervals
  - Bayesian Inference: Credible Intervals

Frequentist and Bayesian Inference

"Bayesians assume that they have some prior knowledge about the world, and they use the sample to update this knowledge. By contrast, frequentists attempt to make precise statements about the world without having any prior knowledge in hand."

- Given sufficient data, the difference between Bayesian and frequentist solutions becomes neglible, and similar strategies can generally be employed for both approaches, but there are some differences to be aware of.

Confidence Intervals

- For population means and proportions, the confidence interval is the sample mean or proportion plus or minus some constant times the standard error.
- Correct Interpretation: "We are XX% confident that the population parameter falls between the lower and upper bound of the interval."
  - It is NOT true that there is a XX% chance or probability of the parameter falling in the interval.
  - It is NOT true that the lower bound is the minimum possible value of the parameter and the upper bound is the maximum possible value.

Credible Intervals

- Bayesians calculate posterior distributions and credible intervals.
- The Bayesian posterior distribution tells us how likely specific parameter estimates are given the input data.
- The credible interval indicates a range of values in which the parameter value is expected with a given probability, as calculated from the posterior distribution.
- The central goal of Bayesian estimation is to obtain the posterior distribution. **Therefore, Bayesians commonly visualize the entire distribution rather than simplifying it into a credible interval.** In terms of data visualization, therefore, all the approaches to visualizing distributions discussed previously are applicable.
  - Histograms, density plots, boxplots, violins, ridgeline plots, etc.

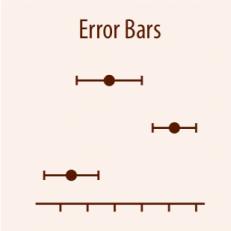
Visualizing Uncertainty

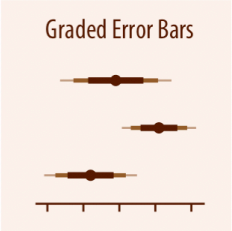
"When we see a data point drawn in a specific location, we tend to interpret it as a precise representation of the true data value. It is difficult to conceive that a data point could actually lie somewhere it hasn't been drawn. Yet this scenario is ubiquitous in data visualization. Nearly every data 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](https://clauswilke.com/dataviz/visualizing-uncertainty.html)

Visualizing Uncertainty in Means and Proportions 1

- **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.





Error Bars 1

- Error bars are convenient because they allow us to show many estimates with their uncertainties all at once.
- Therefore, they are commonly used in scientific publications, where the primary goal is usually to convey a large amount of information to an expert audience.
- Error bars can also be easily combined with other types of plots! For example, bar charts, scatters, etc.

Bar Charts with Error

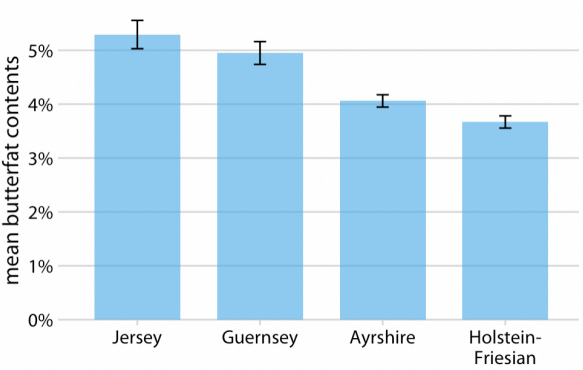


Figure 16.10: Mean butterfat contents in the milk of four cattle breeds. Error bars indicate +/- one standard error of the mean. Visualizations of this type are frequently seen in the scientific literature. While they are technically correct, they represent neither the variation within each category nor the uncertainty of the sample means particularly well. See Figure 7.11 for the variation in butterfat contents within individual breeds. Data Source: Canadian Record of Performance for Purebred Dairy Cattle

Scatterplots with Error

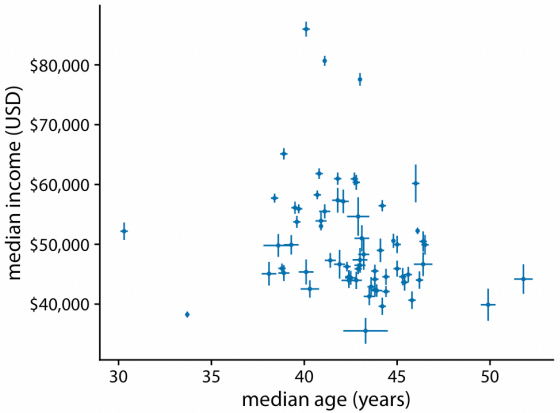
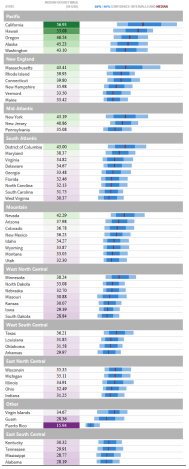


Figure 16.11: Median income versus median age for 67 counties in Pennsylvania. Error bars represent 90% confidence intervals. Data source: 2015 Five-Year American Community Survey

Graded Error Bars

- If we show simple error bars to a group of people, chances are at least some of them would perceive the error bars deterministically.
  - For example, representing minimum and maximum of the data.
  - Alternatively, they might think the error bars delineate the range of possible parameter estimates, i.e., the estimate could never fall outside the error bars.
- These types of misperception are called deterministic construal errors. The more we can minimize the risk of deterministic construal error, the better our visualization of uncertainty.
- Graded error bars show multiple different confidence intervals at the same time, using darker colors and thicker lines for the intervals representing lower confidence levels.
- The grading helps the reader perceive that there is a range of different possibilities.

Graded Error Bars: In the Wild



Error Bars 2

- For confidence intervals for means and proportions, there is always a level of confidence associated with the construction of the interval.
  - Common values: 80%, 90%, 95%, 99%
  - In general, the higher the confidence, the wider the interval.
- **Whenever you visualize uncertainty with error bars, you must specify what quantity and/or confidence level the error bars represent.**

Significant Differences 1

- One common question we are trying to answer is "are the means for various groups significantly different?"
- The word "significant" here is a technical term used by statisticians. We call a difference significant if with some level of confidence we can reject the assumption that the observed difference was caused by random sampling.

Significant Differences 2

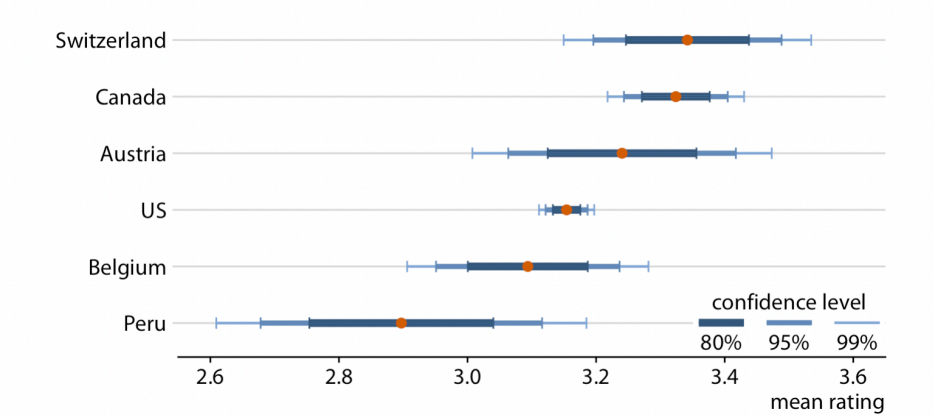


Figure 16.7: Mean chocolate flavor ratings and associated confidence intervals for chocolate bars from manufacturers in six different countries. Data source: Brady Brelinski, Manhattan Chocolate Society

Significant Differences 3

- When comparing two groups, we have to take uncertainty from both means into account.
- Statistics textbooks and online tutorials sometimes publish rules of thumb of how to judge significance from the extent to which error bars do or don't overlap. **However, these rules of thumb are not reliable and should be avoided.**
- The correct way to assess whether there are differences in mean rating is to calculate confidence intervals for the differences. If those confidence intervals exclude zero, then we know the difference is significant at the respective confidence level.

Significant Differences 4

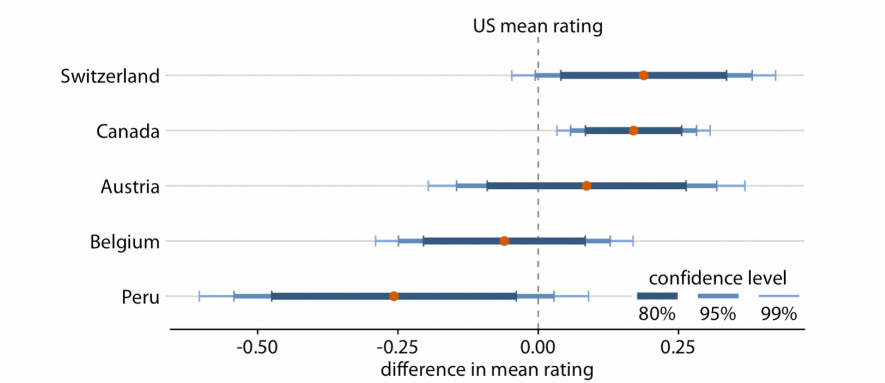


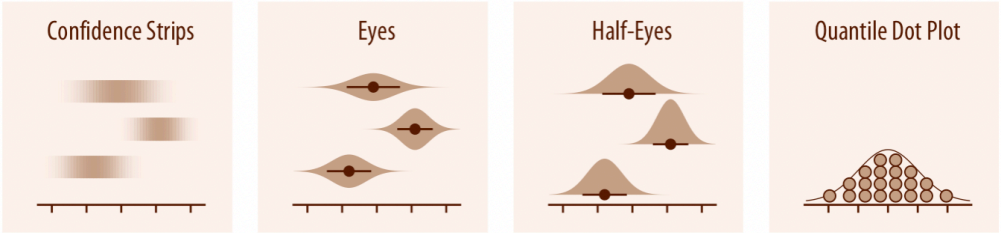
Figure 16.8: Mean chocolate flavor ratings for manufacturers from five different countries, relative to the mean rating of U.S. chocolate bars. Canadian chocolate bars are significantly higher rated that U.S. bars. For the other four countries there is no significant difference in mean rating to the U.S. at the 95% confidence level. Confidence levels have been adjusted for multiple comparisons using Dunnett's method. Data source: Brady Brelinski, Manhattan Chocolate Society

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

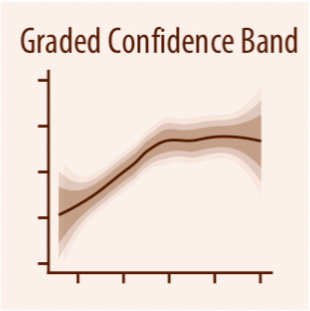
- As an alternative to error bars we could draw confidence strips that gradually fade into nothing.
- Confidence strips better convey how probable different values are, but they are difficult to read.
- We would have to visually integrate the different shadings of color to determine where a specific 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.

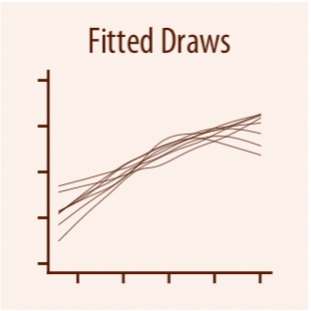
Visualizing Uncertainty in Trends 1

- 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.

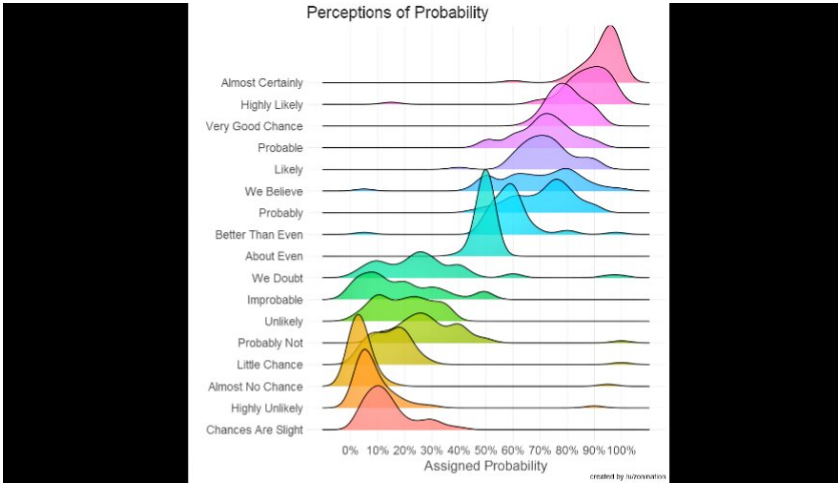




Probability 1

- One tool for dealing with uncertainty is probability–it's really easy to understand that there is a lot of uncertainty with regards to future events.
- There are different ways of thinking about probability, but a relatively simple way is to think about counting the number of times an event occurs relative to the number of times an event could occur.
- In general, mathematicians like to think about events in the long run, a.k.a., what would happen if the number of times the event could occur as infinite.
  - If we flipped a coin an infinite number of times, we would expect to see "Heads" in 50% of the flips.
  - If we rolled a dice an infinite number of times, we would expect to see a 4 in one out of every six rolls.

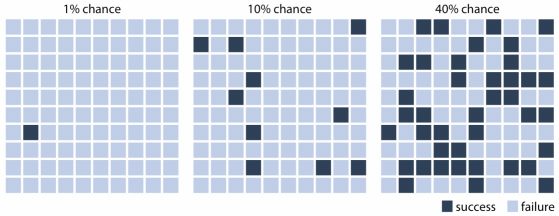
Probability Perception



@panlepan

Discrete Outcome Visualizations

- We can make the concept of probability tangible by creating a graph that emphasizes both the frequency aspect and the unpredictability of a random trial, for example by drawing squares of different colors in a random arrangement.
- This style of visualization, where we show specific potential outcomes, is called a discrete outcome visualization, and the act of visualizing a probability as a frequency is called frequency framing.



Election Forecast

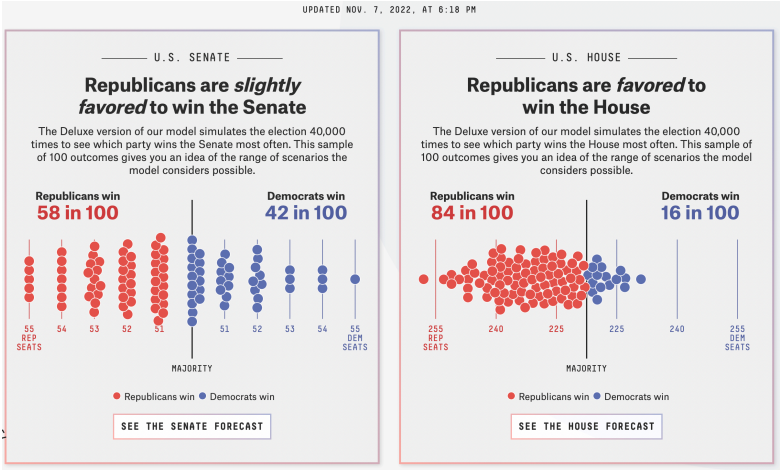
DATA 227

Vocabulary Review

Variability

Uncertainty

Probability



Five Thirty Eight Election Forecast

Sources

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