# Uncertainty Visualization

Understanding and Communicating Uncertainty in Data Visualization

# Introduction to Uncertainty

- Uncertainty is inherent in most data sets.
- Visualizing uncertainty is crucial for accurate interpretation.
- Challenges: Our brains tend to assume data points are exact.
- Different approaches cater to different audiences (experts vs. laypeople).

# Common Techniques for Visualizing Uncertainty

- **Error Bars:** Indicate uncertainty for individual points.
- **Confidence Bands:** Show uncertainty across a range.
- **Frequency Framing:** Uses repeated scenarios to illustrate probability.
- **Animation:** Depicts uncertainty dynamically over time.

# Framing Probabilities as Frequencies

- Probability is a measure of uncertainty.
- Best understood through frequency framing.
- Example: Coin flips, rolling dice.
- Importance: Translating probability values into real-world experience.

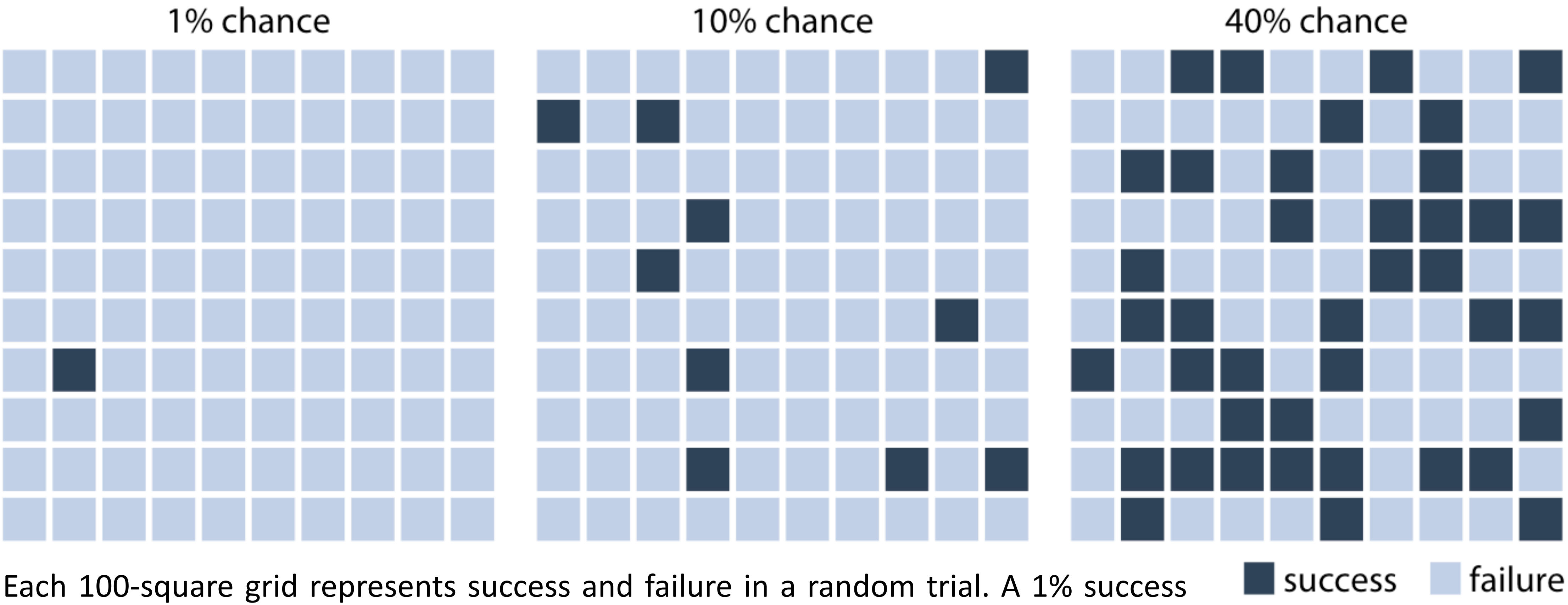
# Discrete Outcome Visualization

- Uses frequency framing to illustrate probability.
- Example: Grid of 100 squares with success/failure outcomes.
- Helps viewers intuitively understand probability through visual randomness.

# Example of Frequency Framing

- Visualizing probabilities (1%, 10%, 40%) using discrete squares.
- Helps illustrate likelihood and randomness.
- Overcomes limitations of single-number probability representation.

# Visualizing probability as frequency




Each 100-square grid represents success and failure in a random trial. A 1% success rate has 1 dark square, 10% has 10, and 40% has 40. Random placement of dark squares emphasizes outcome uncertainty in a single trial.

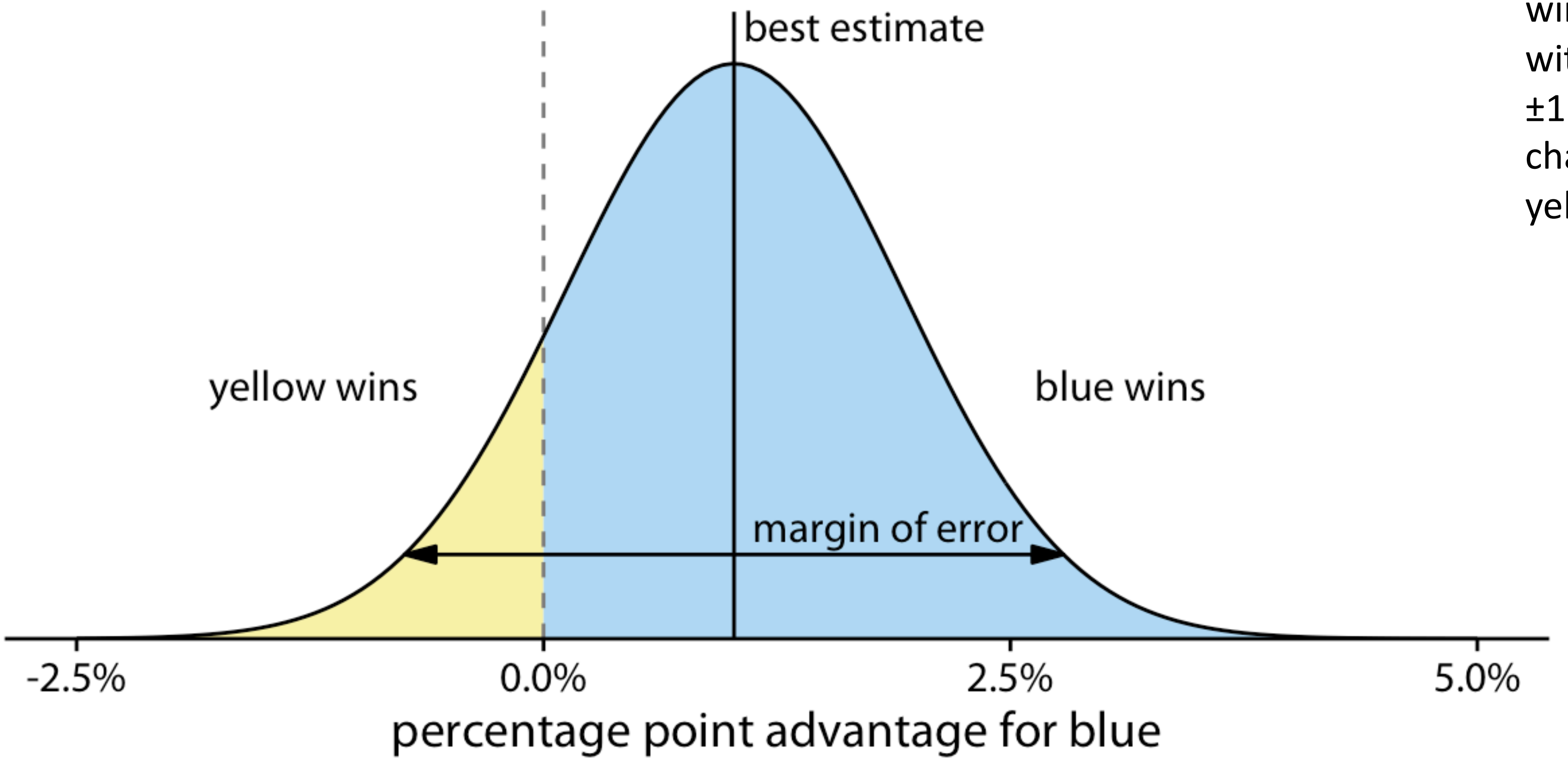
# Probability Distributions

- Used when outcomes are continuous rather than discrete.
- Example: Election predictions with a margin of error.
- Shows a range of possible outcomes, not just a single number.

# Election Prediction Visualization

- Blue party predicted to win by 1% with a margin of error of 1.76%.
- **Shaded Areas:** 87.1% chance blue wins, 12.9% chance yellow wins.
- Challenges: People tend to focus on the most likely outcome rather than uncertainty.

# Hypothetical election prediction




The blue party is estimated to win by one percentage point, with a 95% margin of error of ±1.76 points. Blue has an 87% chance of winning, while yellow has 13%.

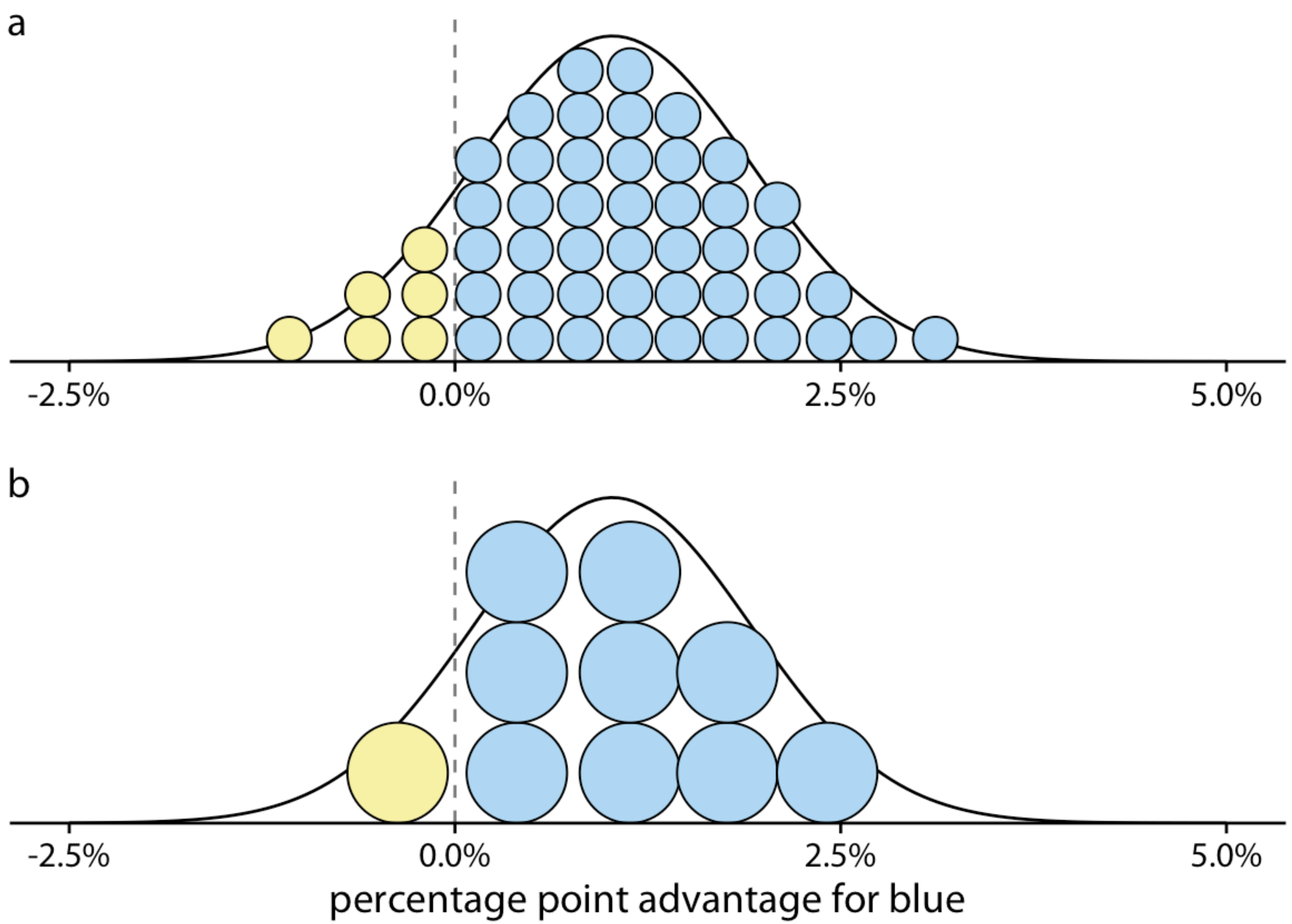
# Comparing Probability Visualizations

- Discrete outcome visualization (e.g., frequency framing) vs. smooth probability distributions.
- Research shows frequency framing is more intuitive for non-expert audiences.

# Quantile Dot Plots

- Combines discrete outcomes with continuous probability distributions.
- Example: Representing election outcomes with 50 vs. 10 dots.
- Fewer dots improve readability while maintaining accuracy.

# Quantile dotplot of election outcome




(a) 50 dots (2% each) approximate the distribution, with six yellow dots indicating a 12% chance. (b) 10 dots (10% each) simplify the visualization, with one yellow dot representing a 10% chance. Fewer dots improve readability.

# Statistical Sampling and Uncertainty

- In statistics, we use samples to estimate population parameters.
- **Population:** The complete set of data.
- **Sample:** A subset used to make inferences about the population.
- **Estimates:** Values derived from a sample, including means and standard deviations.
- **Standard error (SE):** Measures the precision of an estimate.

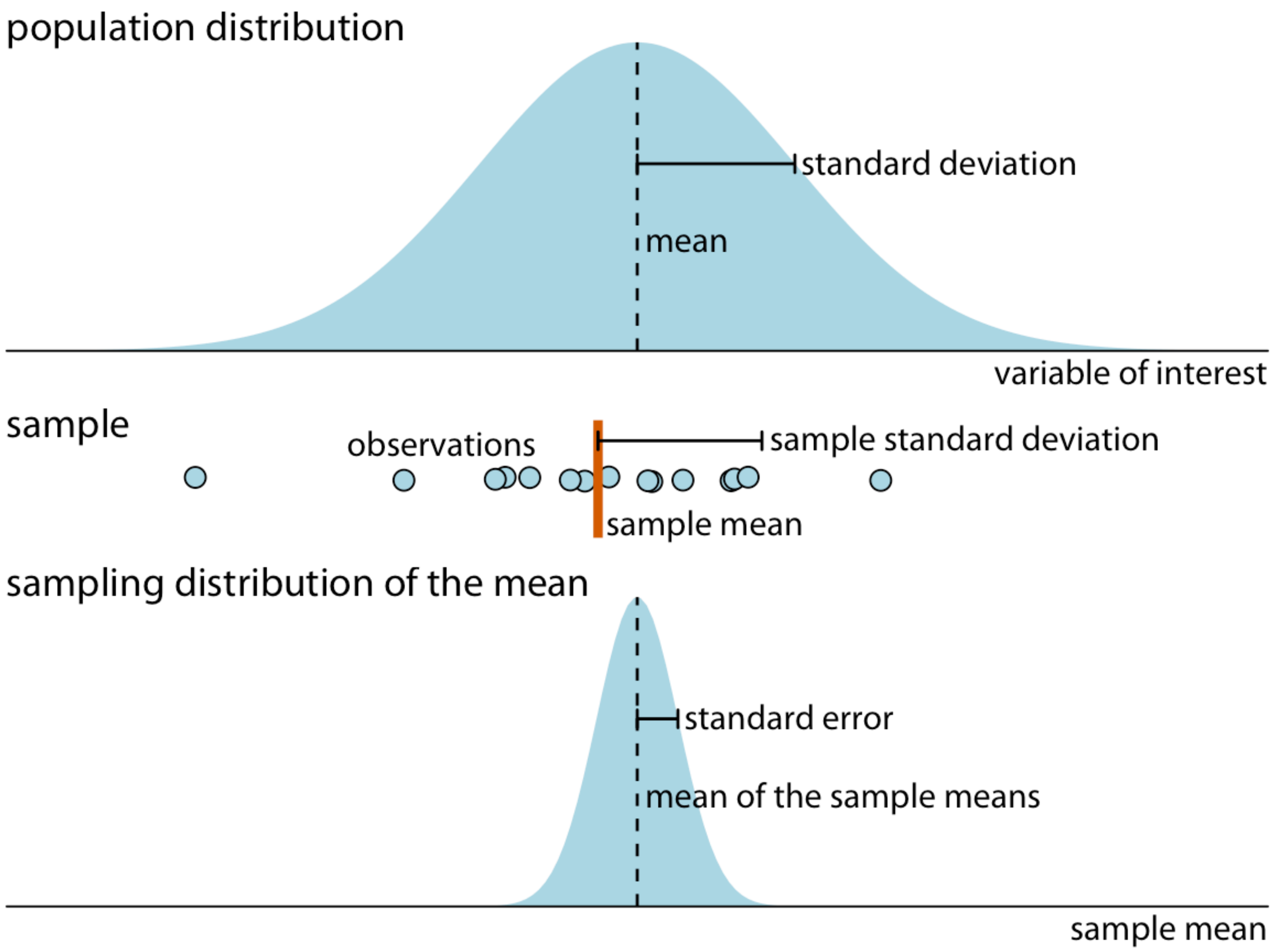
# Point Estimates and Margins of Error

- A point estimate is a single value used to estimate a population parameter.
- The margin of error provides a range within which the true parameter likely falls.
- Example: Election polling estimates with confidence intervals.

# Standard Deviation vs. Standard Error

- **Standard Deviation (SD):** Describes the spread of individual observations in the population.
- **Standard Error (SE):** Describes the precision of a sample estimate (SE = SD / sqrt(sample size)).
- Larger samples reduce SE, increasing confidence in estimates.

# Key concepts of statistical sampling




A population has a true distribution with a mean and standard deviation. Sample means vary from the population mean, forming a sampling distribution. The standard error indicates the precision of the estimate.

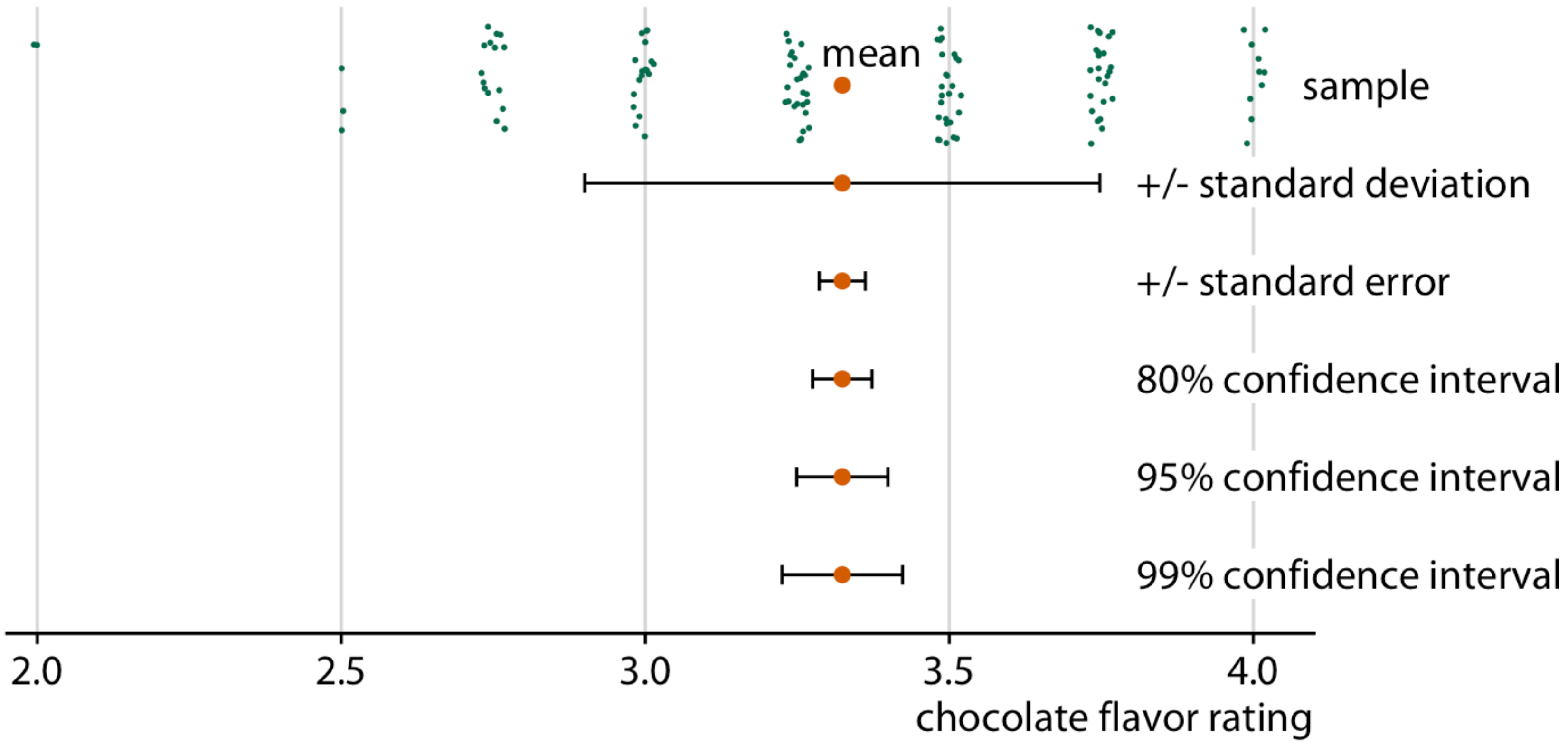
# Frequentist vs. Bayesian Approaches

- **Frequentist:** Estimates uncertainty based solely on sample data (e.g., confidence intervals).
- **Bayesian:** Incorporates prior knowledge to update estimates.
- Both methods use similar visualization techniques for uncertainty.

# Visualizing Uncertainty with Error Bars

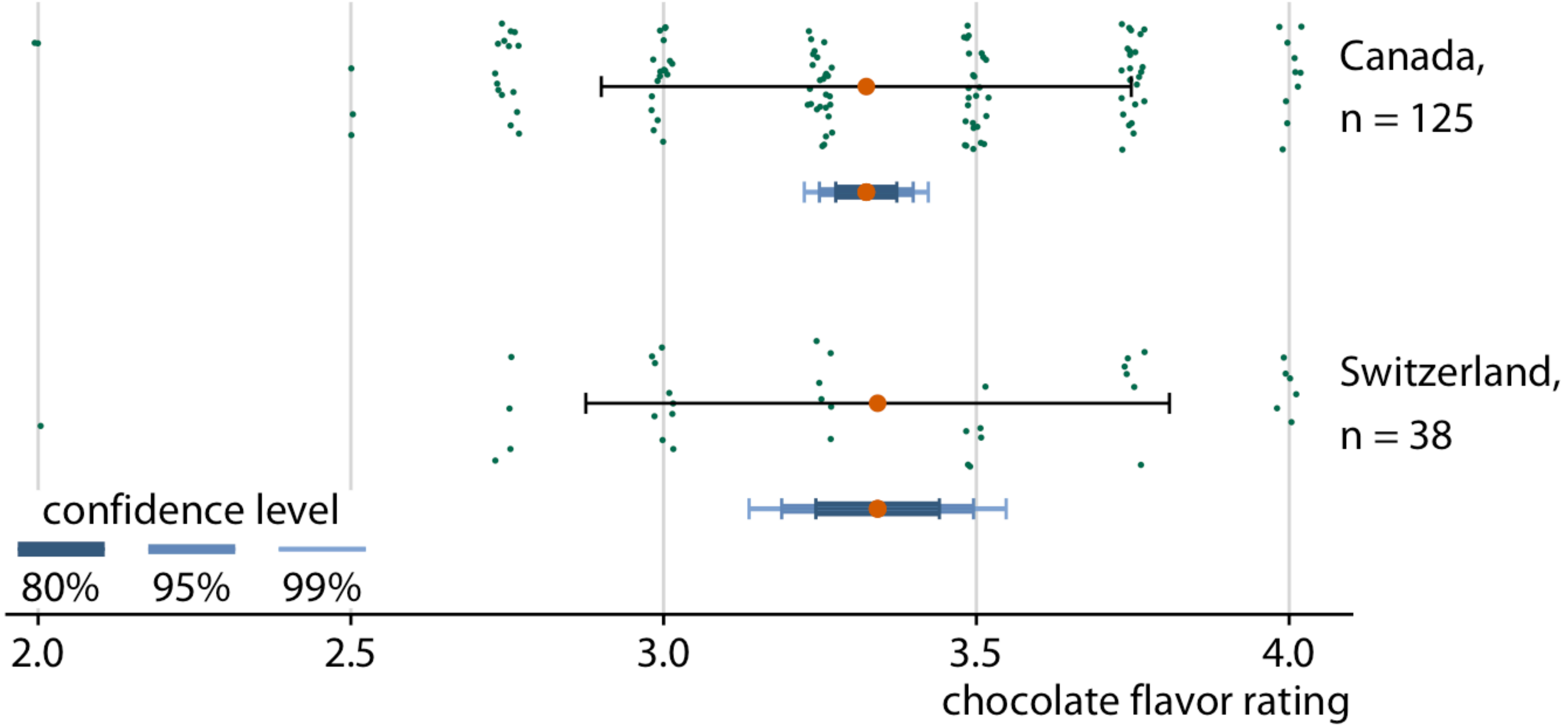
- **Error bars** represent uncertainty in estimates.
- Common types:
  - Standard deviation
  - Standard error
  - Confidence intervals (e.g., 80%, 95%, 99%)
- Example: Chocolate ratings dataset with different error bars.

# Sample statistics for chocolate bar ratings




Relationship between sample, mean, standard deviation, standard error, and confidence intervals in chocolate bar ratings. Green dots represent expert ratings of 125 Canadian chocolate bars (scale: 1 to 5). The orange dot shows the mean. Error bars depict twice the standard deviation, twice the standard error, and 80%, 95%, and 99% confidence intervals.

# Confidence intervals widen with smaller samples

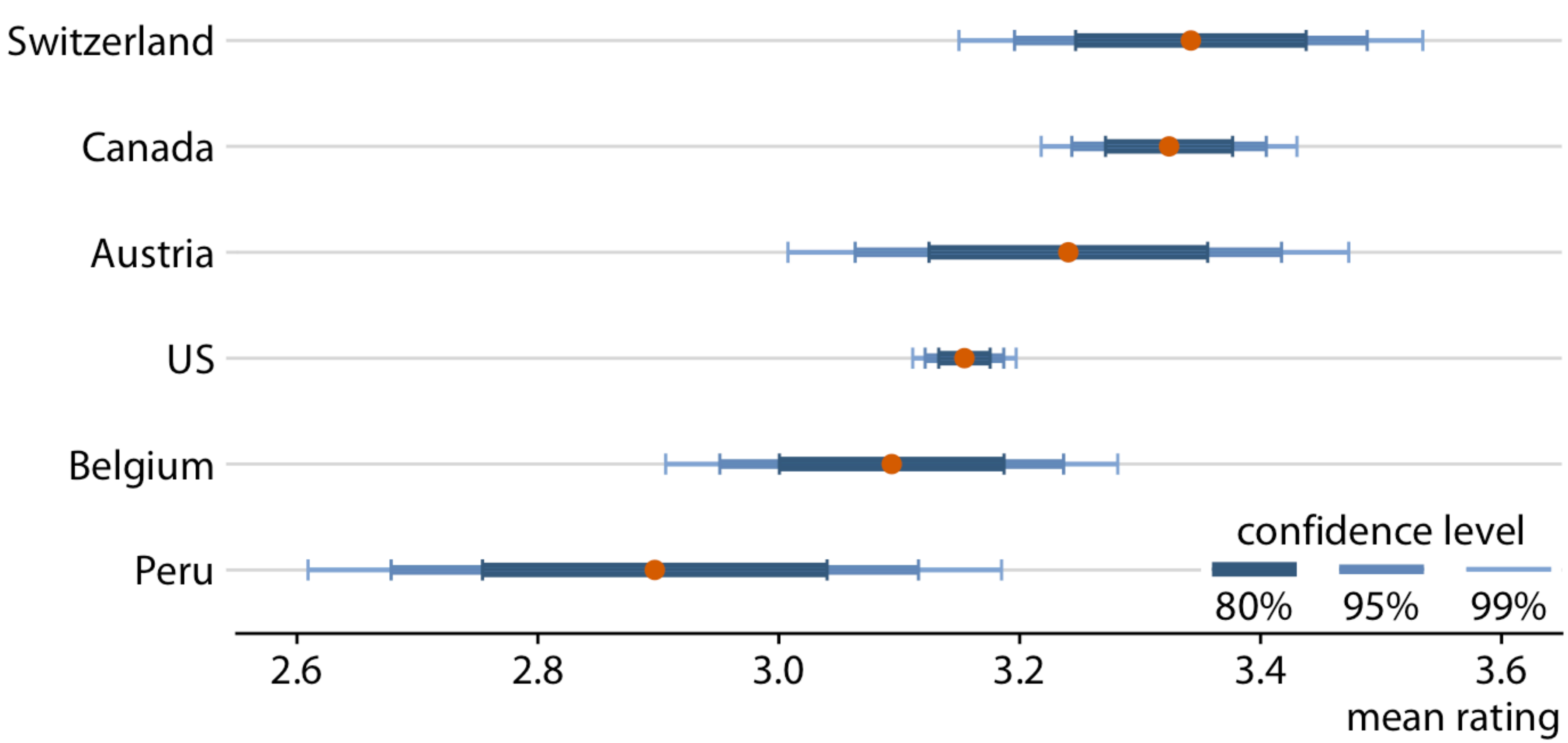



Swiss and Canadian chocolate bars have similar mean ratings and standard deviations (black error bars), but the Swiss sample is smaller, resulting in wider confidence intervals.

# Confidence Intervals and Their Interpretation

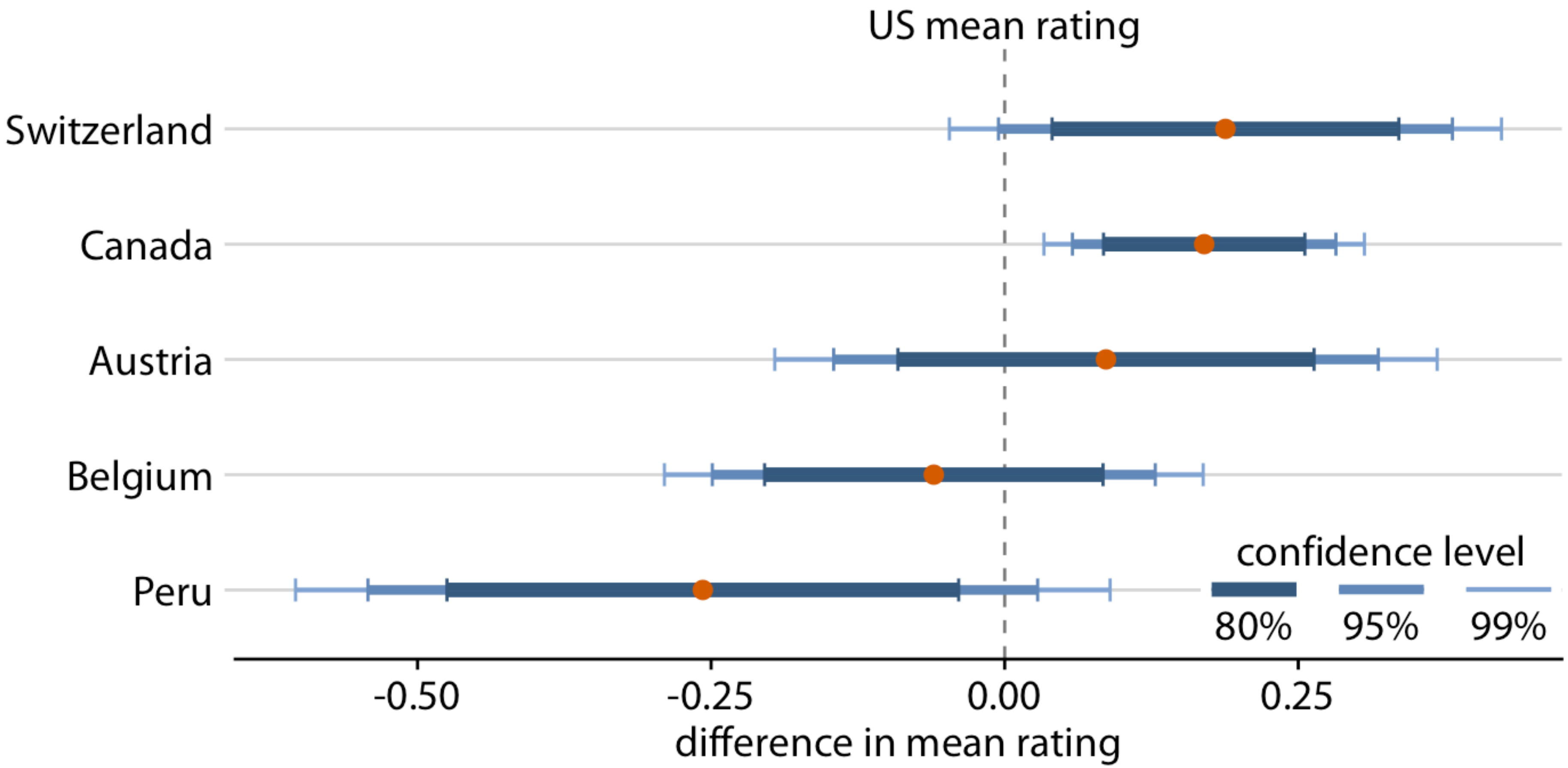
- Confidence intervals (CIs) show a range of plausible values for an estimate.
- **Example:** 95% CI means that if we repeated sampling, 95% of intervals would contain the true mean.
- Larger sample sizes result in narrower CIs.

# Chocolate Flavor Ratings




Mean chocolate flavor ratings and associated confidence intervals for chocolate bars from manufacturers in six different countries.

# Mean Chocolate Ratings Relative to U.S. Bars

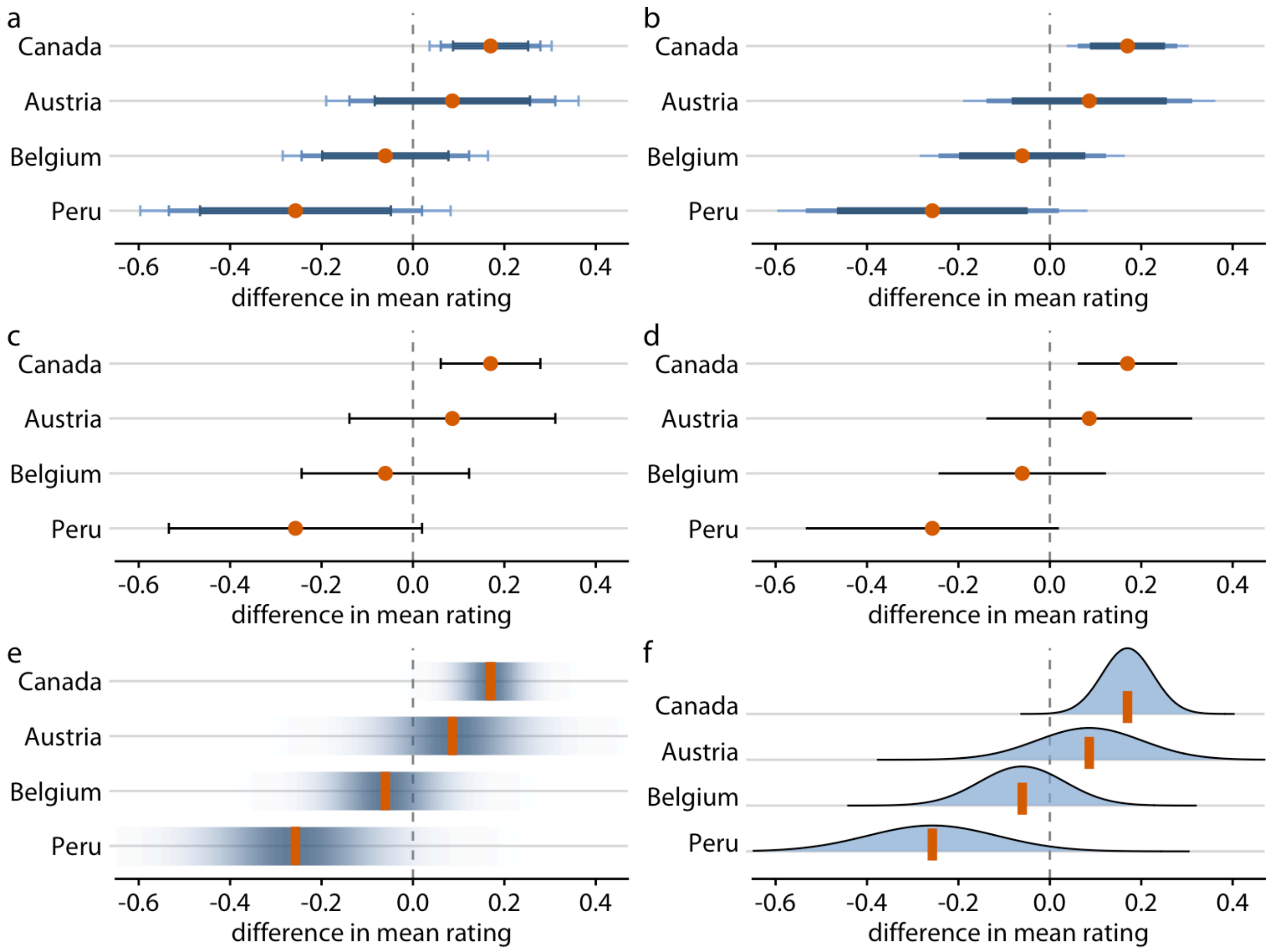



Mean chocolate flavor ratings for five countries, relative to U.S. bars. Canadian chocolate is rated significantly higher; no significant difference for others.

# Overlapping Confidence Intervals and Significance

- Misconception: Overlapping CIs always imply non-significant differences.
- Correct approach: Calculate confidence intervals for the difference between means.
- Example: Chocolate ratings comparison across countries.

# Different Uncertainty Approaches




Mean chocolate flavor ratings for four countries, relative to U.S. bars, using different uncertainty visualization approaches: (a) Graded error bars with cap, (b) without cap, (c) single-interval error bars with cap, (d) without cap, (e) confidence strips, (f) confidence distributions.

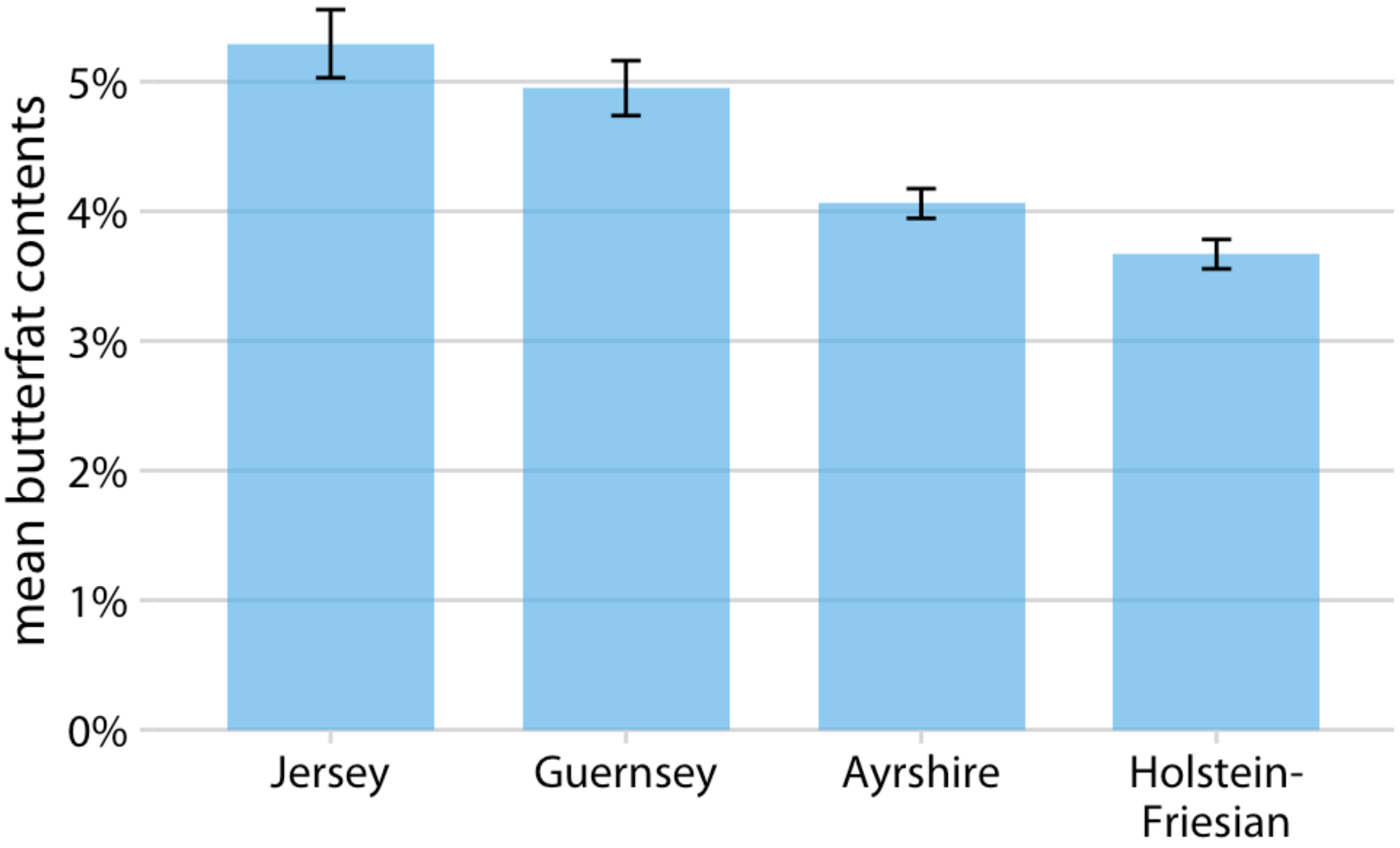
# Alternative Uncertainty Visualizations

- **Graded error bars:** Show multiple confidence levels.
- **Confidence strips:** Gradually fade to indicate uncertainty.
- **Confidence distributions:** Show probability density of estimates.
- **Quantile dotplots:** Offer clearer probability representation.

# Error Bars and Their Importance

- Error bars are a simple and effective way to represent uncertainty in data visualizations.
- They can be added to various types of plots, such as bar charts and scatter plots.
- Commonly used in scientific publications to depict variability and confidence intervals.
- Example: Bar plot with error bars representing mean butterfat content in milk

# Mean butterfat contents in milk

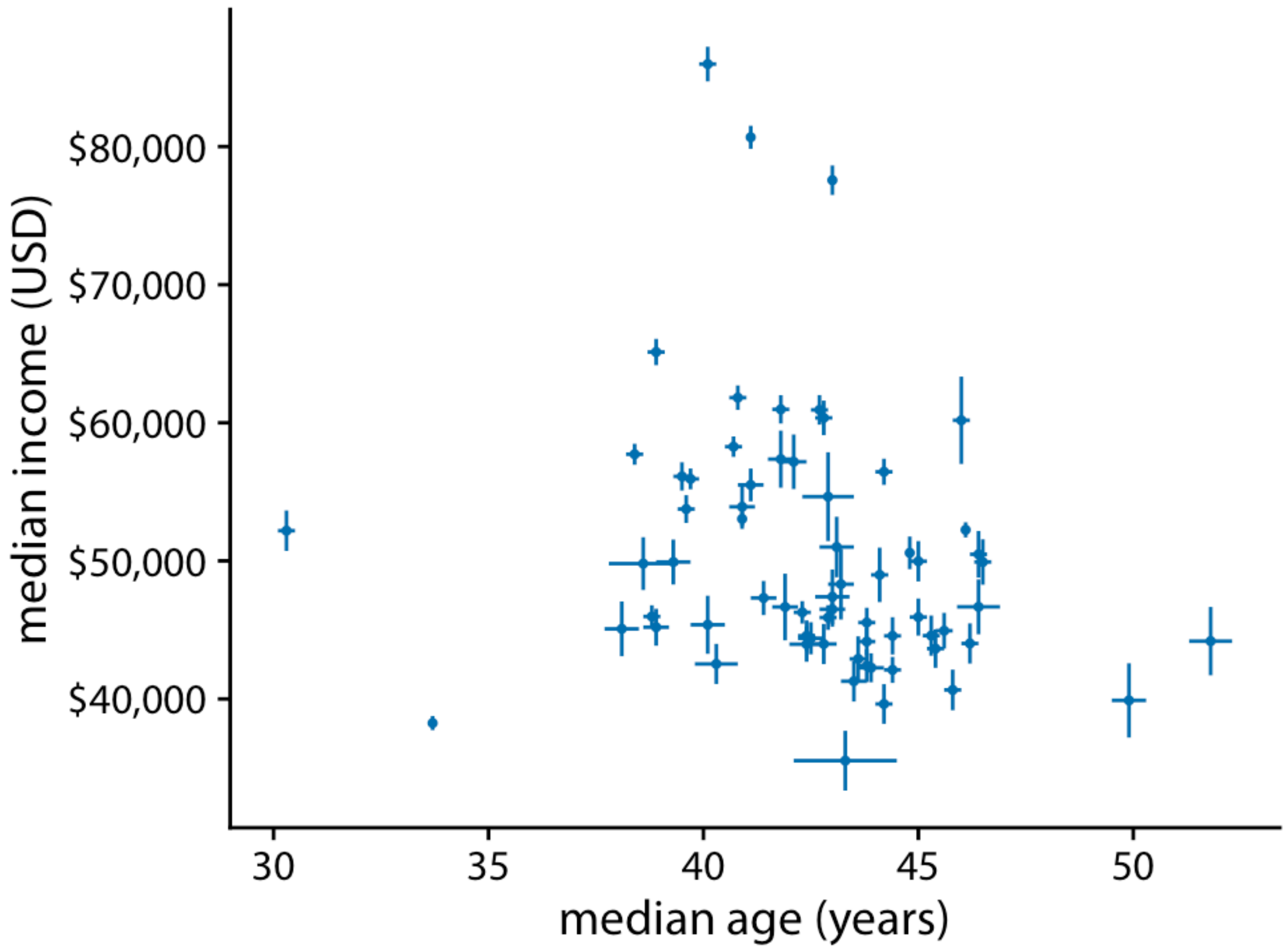



Error bars represent ±1 standard error of the mean. These visualizations, while technically correct, do not fully represent variation within categories or sample mean uncertainty.

# Error Bars in Scatter Plots

- Error bars can be drawn along both x and y axes in scatter plots.
- Useful for showing confidence intervals in data distributions.
- Example: Median income vs. median age for 67 counties in Pennsylvania

# Median income versus median age

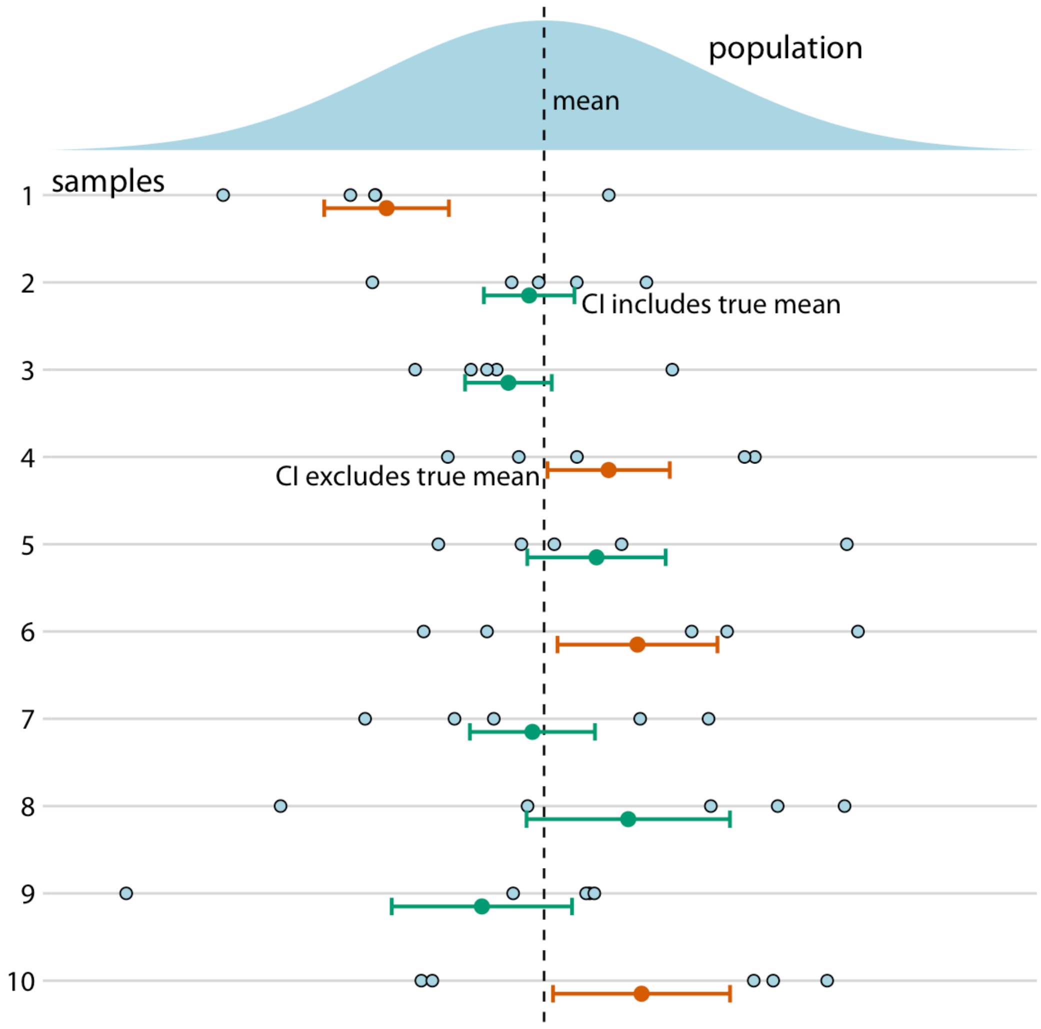



Median income versus median age for 67 counties in Pennsylvania. Error bars represent 90% confidence intervals.

# Frequentist vs. Bayesian Uncertainty Interpretation

- Frequentists use confidence intervals, while Bayesians use posterior distributions and credible intervals.
- A Bayesian credible interval provides a probability range where the true parameter is expected to lie.
- A frequentist confidence interval represents a range where the true parameter is likely not located under repeated sampling

# Frequency of confidence interval

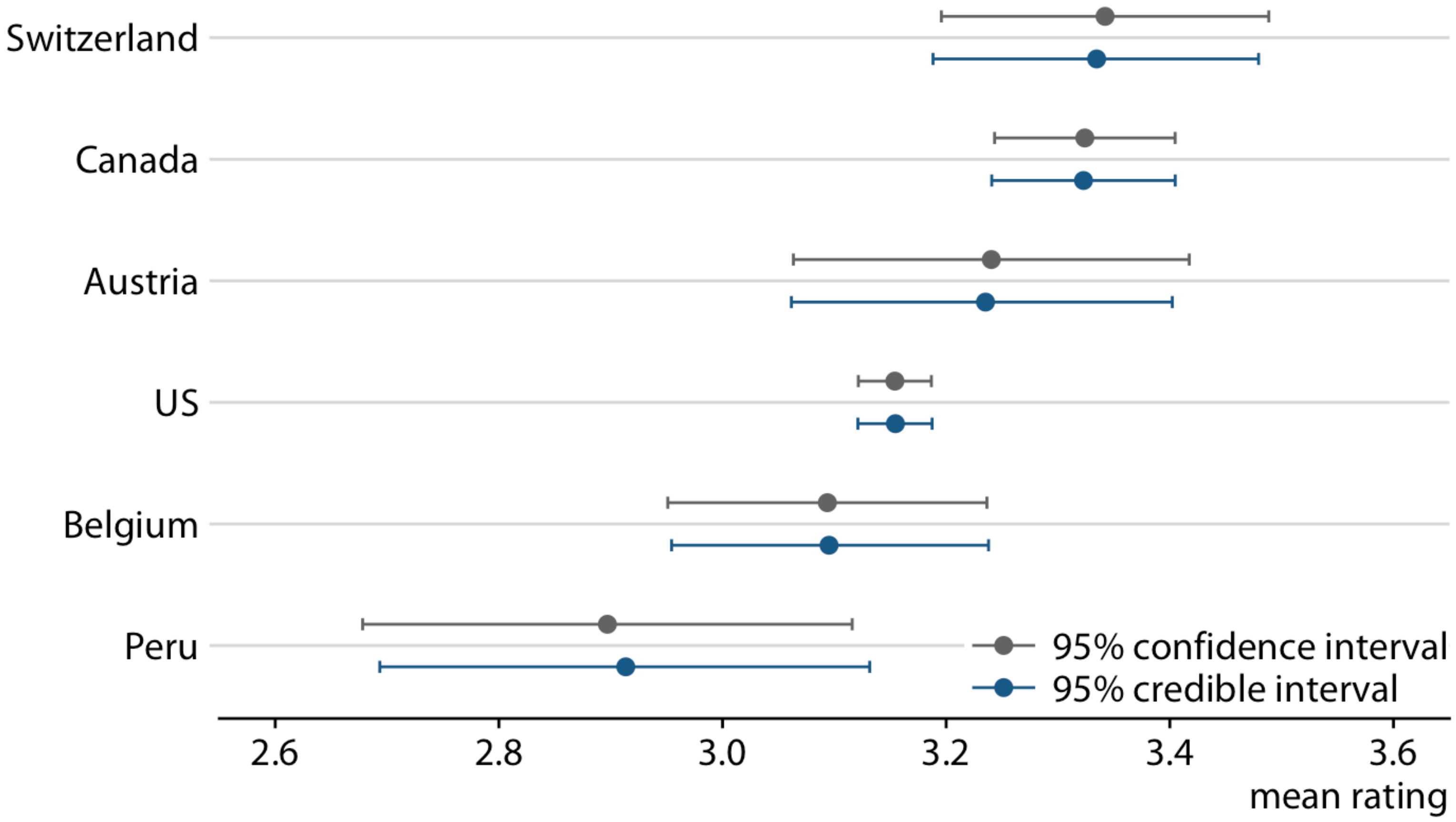



In repeated sampling, 68% confidence intervals (sample mean ± standard error) include the true mean about 68% of the time.

# Frequentist vs. Bayesian Estimates Comparison

- Confidence intervals and credible intervals often yield similar results but have conceptual differences.
- Bayesian estimates exhibit shrinkage, adjusting extreme values toward the mean.
- Example: Mean chocolate ratings comparison between frequentist and Bayesian approaches

# Frequentist vs Bayesian Interval Estimates for Chocolate Ratings

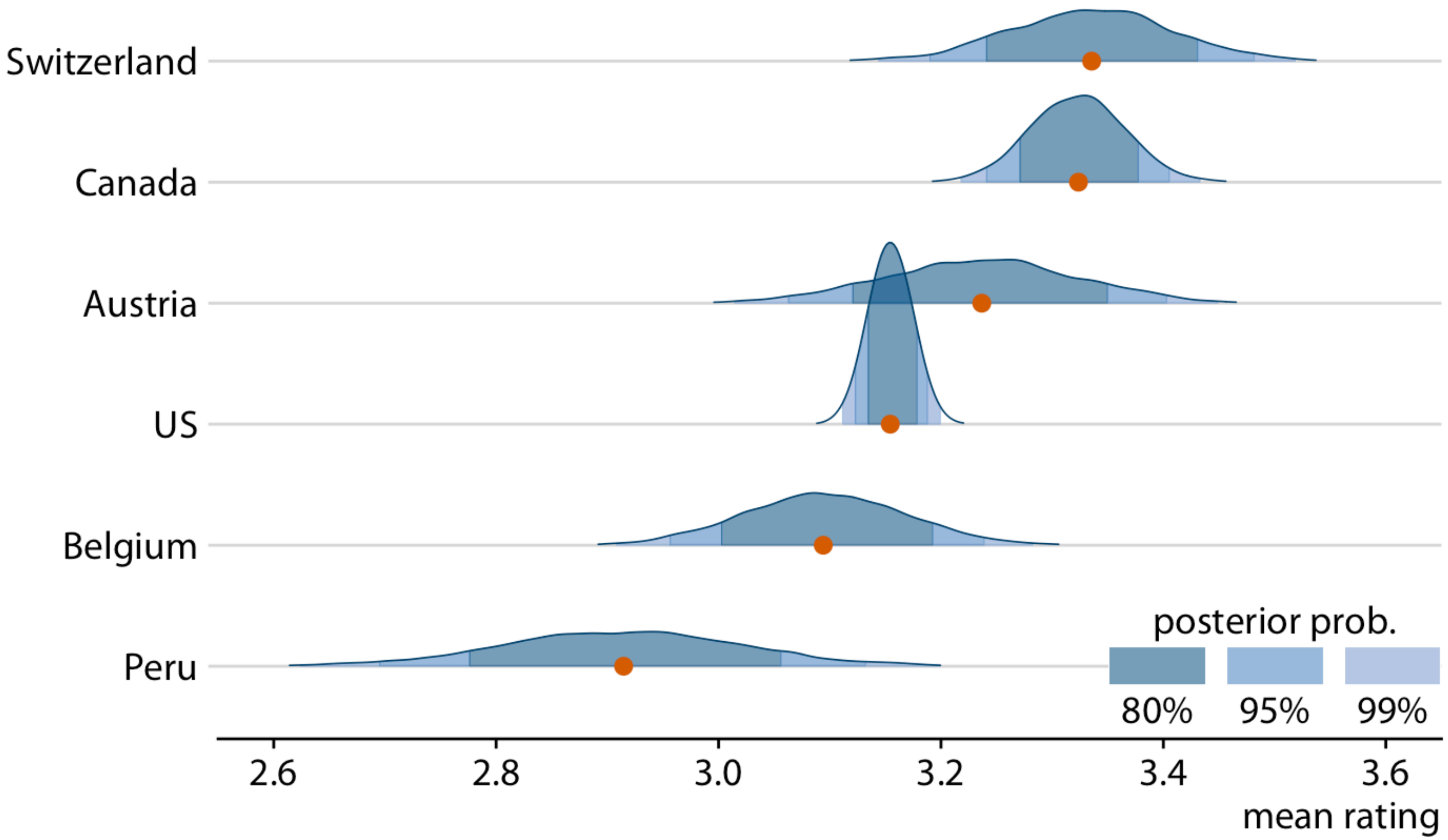



Comparison of frequentist confidence intervals and Bayesian credible intervals for mean chocolate ratings. Both approaches yield similar results, with Bayesian estimates showing slight shrinkage, adjusting extreme values towards the overall mean (e.g., Switzerland and Peru).

# Bayesian Posterior Distributions

- Bayesians visualize full posterior distributions rather than just intervals.
- Common visualization techniques: histograms, density plots, violins, and ridgeline plots.
- Example: Ridgeline plot of Bayesian posterior distributions of chocolate ratings

# Bayesian Posterior Distributions of Chocolate Ratings

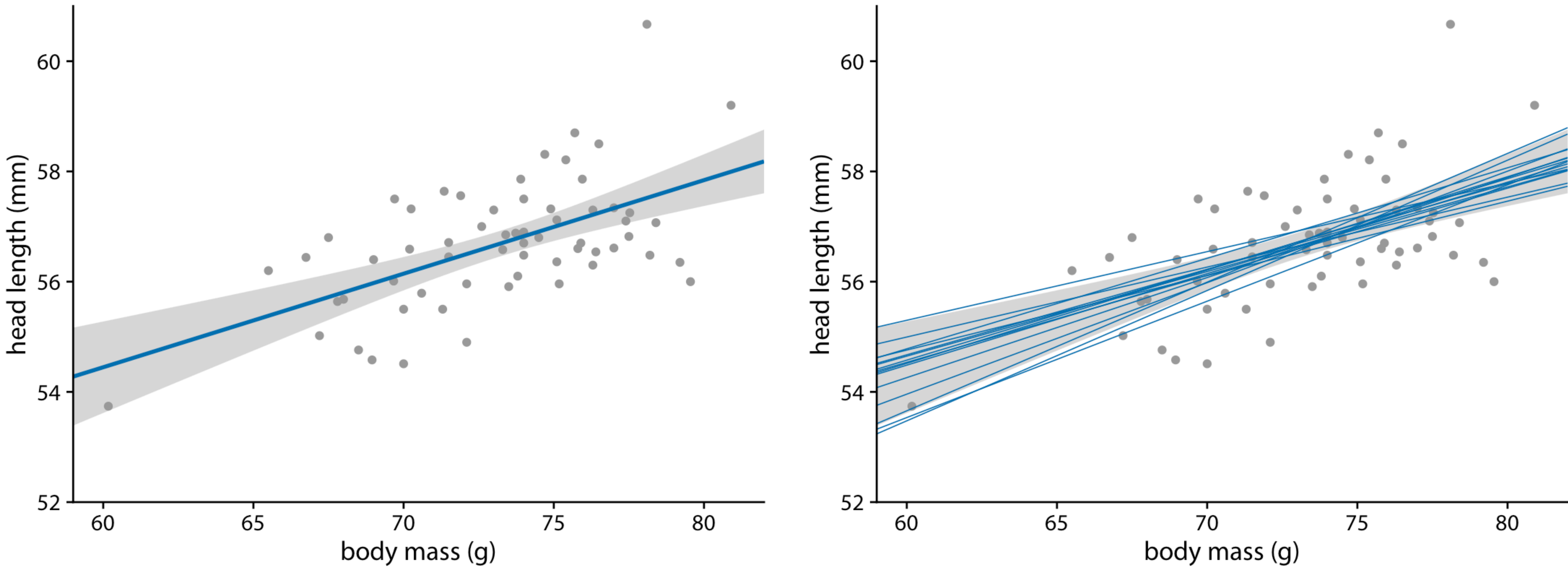



Bayesian posterior distributions of mean chocolate bar ratings shown as a ridgeline plot. Red dots represent the medians, with shading indicating the center 80%, 95%, and 99% of each distribution.

# Uncertainty in Curve Fits

- Trend estimates include uncertainty, often represented as a confidence band.
- Confidence bands provide a range of possible trend lines.
- Example: Head length vs. body mass for male blue jays with confidence bands

# Head Length vs Body Mass for Male Blue Jays

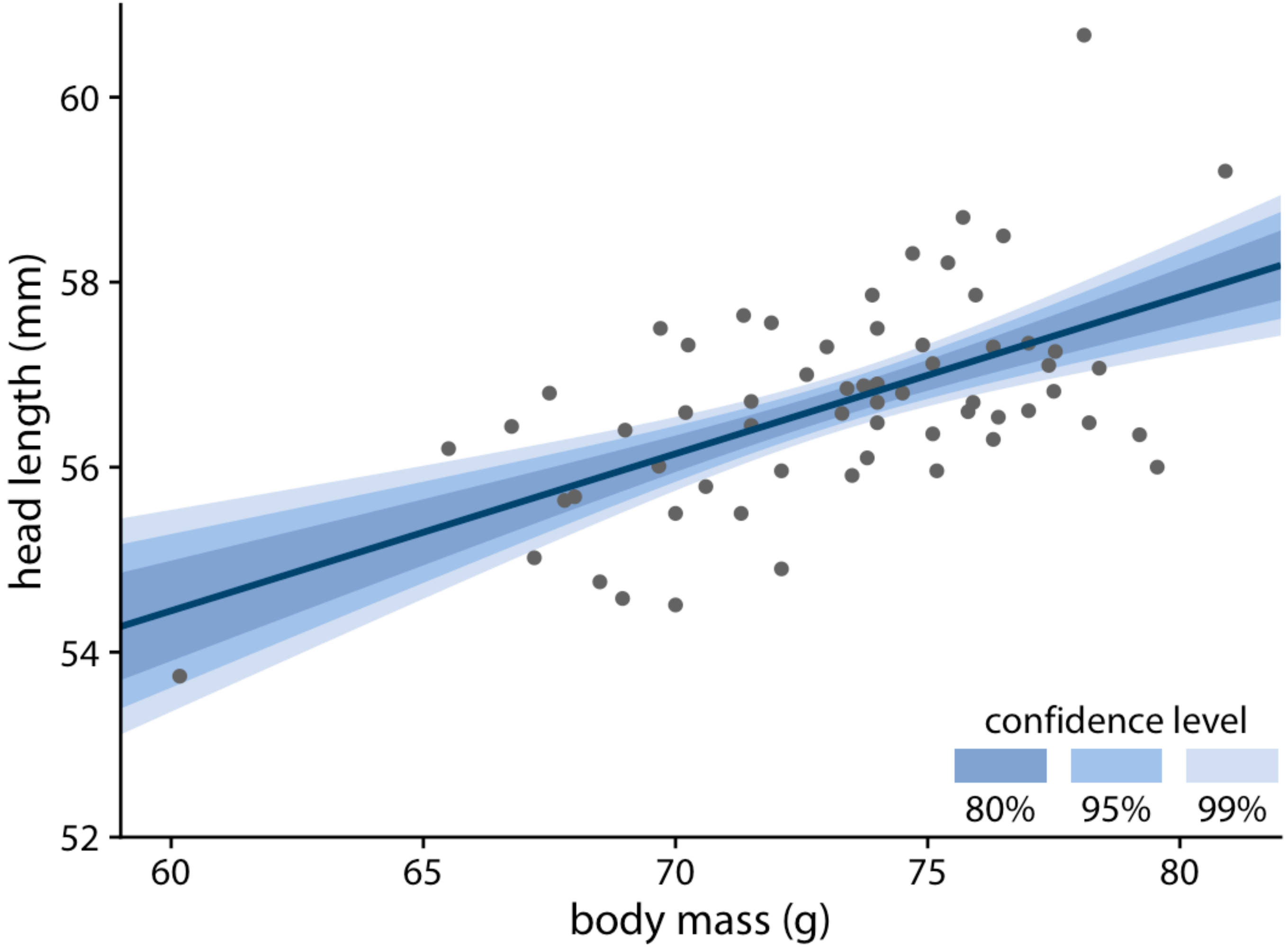



he blue line shows the best linear fit with a 95% confidence band, while the gray band represents uncertainty. In contrast to Figure 16.15, blue lines here represent alternative fits randomly drawn from the posterior distribution.

# Graded Confidence Bands

- Graded confidence bands highlight multiple levels of confidence simultaneously.
- Useful for conveying a more nuanced sense of uncertainty.
- Example: Head length vs. body mass for male blue jays with graded confidence bands

# Head Length vs Body Mass for Blue Jays with Confidence Bands

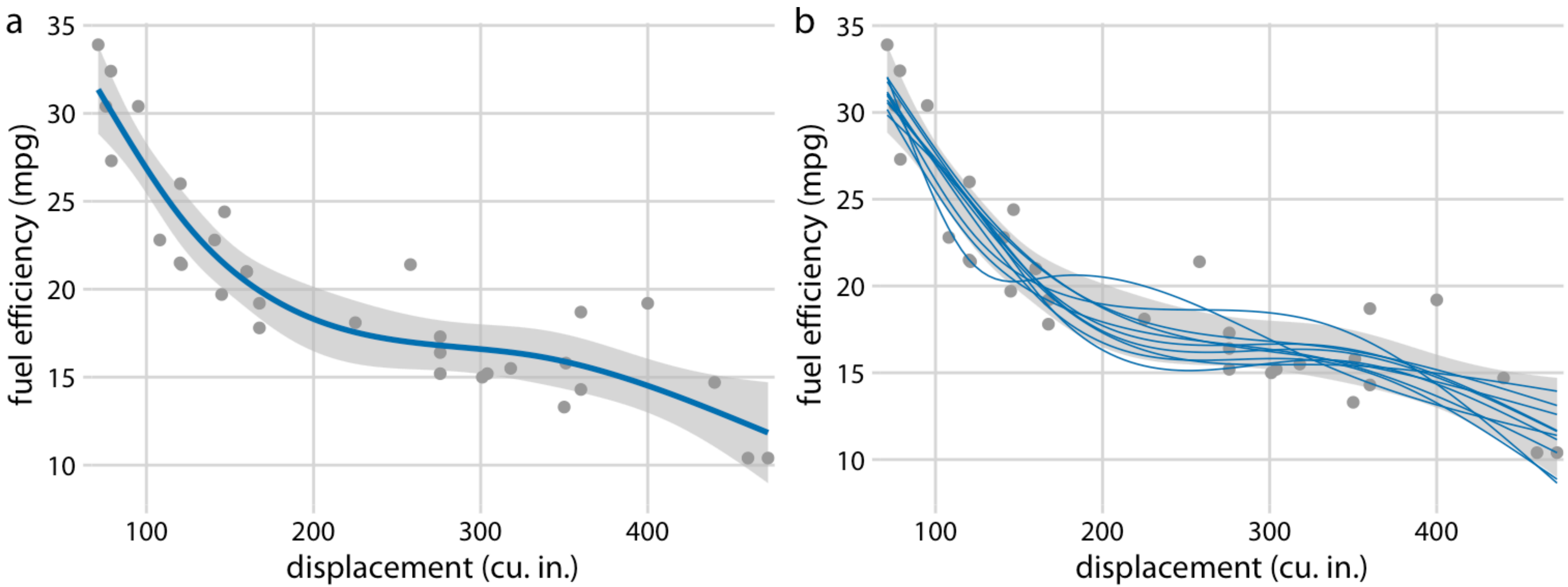



Head length versus body mass for male blue jays. As in the case of error bars, we can draw graded confidence bands to highlight the uncertainty in the estimate.

# Confidence Bands for Non-Linear Fits

- Non-linear confidence bands can be harder to interpret due to curve shape variability.
- Example: Fuel efficiency vs. displacement with cubic regression spline

# Fuel Efficiency vs Displacement with Fit and Alternatives

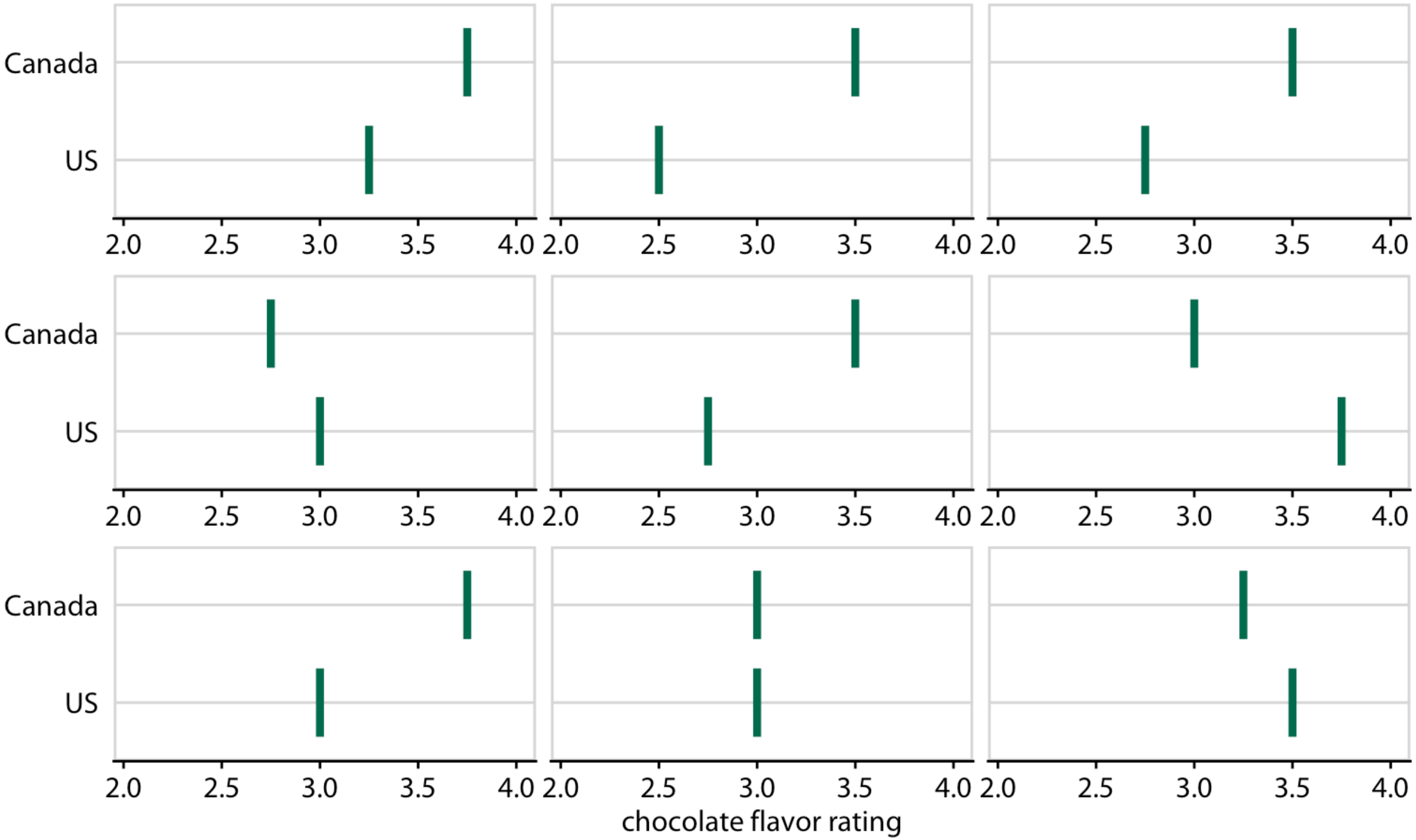



Fuel efficiency vs displacement for 32 cars (1973–74 models). Dots represent individual cars, and smooth lines are cubic regression splines with 5 knots. (a) Best fit spline with confidence band. (b) Alternative fits from the posterior distribution.

# Hypothetical Outcome Plots (HOPs)

- Static uncertainty visualizations can be misinterpreted as deterministic features.
- HOPs use animation to cycle through different likely outcomes.
- Example: Chocolate bar ratings comparison using HOPs

# Hypothetical Outcome Plot for Chocolate Bar Ratings

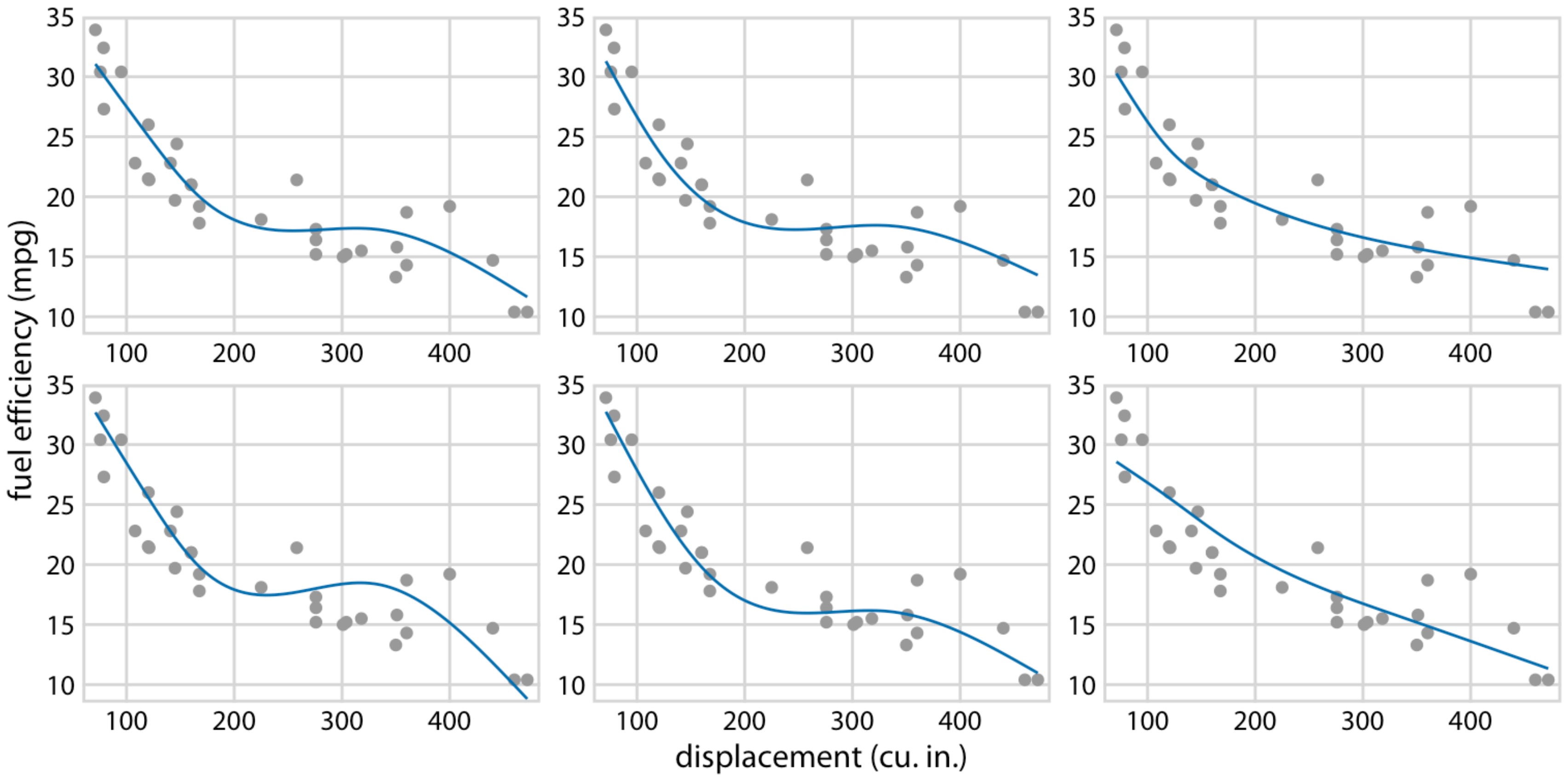



Schematic of a hypothetical outcome plot for chocolate bar ratings. Each green bar represents a rating for one bar, with panels comparing randomly chosen Canadian and U.S. bars. In a real plot, the panels would cycle instead of being displayed side-by-side.

# HOPs for Curve Fits

- HOPs can highlight individual trend lines one at a time to improve interpretability.
- Example: Fuel efficiency vs. displacement shown as a HOP

# Hypothetical Outcome Plot for Fuel Efficiency vs Displacement




Schematic of a hypothetical outcome plot for fuel efficiency vs displacement. Dots represent individual cars, and smooth lines are cubic regression splines with 5 knots. Each panel shows alternative fits from the posterior distribution. In a real plot, the panels would cycle instead of being side-by-side.

# Best Practices in Visualizing Uncertainty

- Balance accuracy with ease of interpretation.
- Always specify what error bars represent (e.g., SD, SE, CI).
- Use graded error bars to avoid deterministic construal errors.
- Frequentist and Bayesian approaches differ in interpretation but often yield similar results.
- Graded confidence bands and HOPs provide additional ways to communicate uncertainty visually.
- Animated visualizations enhance audience understanding and engagement.

# Summary

- Uncertainty is a fundamental part of data analysis.
- Various visualization techniques help convey uncertainty effectively.
- Choosing the right approach improves data communication.
- **Final Thought:** A visualization that is mathematically correct but poorly understood is ineffective.