CHAPTER 5

Visualizing Uncertainty

When we create data visualizations, we usually show each data point at one exact position, like a dot on a graph. But in real life, data is rarely perfectly precise. There is almost always some uncertainty: the real value could be slightly higher, lower, or somewhere else nearby.

However, when people look at a graph, they often assume the dot is the exact value. They don't naturally think about the fact that the true value might be different.

Example: Measuring the Average Height of Students

Suppose we measure the average height of students in a class. Our calculation gives: *Average height = 160 cm*. This measurement process isn't perfect, maybe the measuring tool isn't exact, or the sample of students is small. So, the true average might be a little higher or lower.

This is why showing uncertainty is important. If we don't show it, the audience may misunderstand how reliable or precise the data actually is.

To help communicate uncertainty, we often use tools like:

- **Error bars**: small lines extending above and below (or left and right of) a point to show the possible range.

  Example: **Measuring the Average Height of Students**

  If we plot the average height as a single point on a graph, we could add an error bar like: *160 cm ± 3 cm*. This means the true average could be anywhere from 157 cm to 163 cm. Visually, a dot at 160 cm with a vertical line going up to 163 and down to 157.
- **Confidence bands**: shaded regions around a curve or line to show where the true values are likely to fall.

  Example: **Measuring the Average Height of Students**

  If we measure the average height every year for 10 years. We plot a line showing how the average height changes. But each year's measurement has uncertainty. So, we add a shaded band around the line that widens or narrows depending on how uncertain the data is. Visually, a curve showing height over time, with a light shaded region around it representing the range where the true line probably lies.

These visual elements help viewers better understand how trustworthy and precise the data is.

5.1. Visualizing the Uncertainty of Point Estimates

In statistics, we try to learn about the world by looking at a small part of it. For example, in an election with many districts, we might want to predict how each district will vote and the overall average outcome. Since we can't ask every citizen, we poll a subset of citizens in some districts.

- The population is the full set of all possible votes.
- The sample is the subset we actually poll.

We're usually interested in certain numbers that describe the population, such as the mean vote or the variation across districts. These numbers are called parameters and are generally unknown. Using a sample, we can make estimates of these parameters. For example, the sample mean estimates the population mean. These estimates are often called point estimates because each can be represented by a single value.

Figure 5.1 illustrates these ideas:

- The variable of interest (e.g., votes) has a population distribution with a population mean and population standard deviation.
- A sample gives a set of observations. The number of observations is the sample size.
- From the sample, we calculate the sample mean and sample standard deviation, which usually differ from the population values.

- If we repeated the sampling many times, the distribution of the sample means is the sampling distribution.
- The width of this distribution, called the standard error, shows how precise our estimate is: a smaller standard error means less uncertainty.

Key point: Larger samples give smaller standard errors and more reliable estimates.

It is important not to confuse standard deviation and standard error. The standard deviation describes the variability within a population, it shows how much individual observations differ from each other. For example, in voting districts, it indicates how district outcomes vary. In contrast, the standard error measures the precision of a parameter estimate. For instance, if we estimate the mean voting outcome across all districts, the standard error tells us how accurately our sample mean reflects the true population mean.

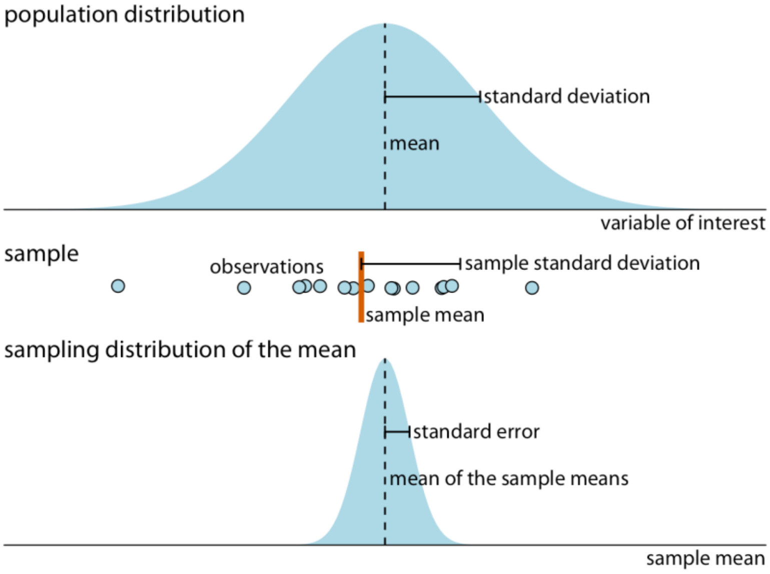

Figure 5.1 Key concepts of statistical sampling. The variable we study has a true distribution in the population, with a population mean and population standard deviation. Any sample we take will have its own sample mean and sample standard deviation, which usually differ from the population values. If we repeated sampling many times, the sample means would form a sampling distribution. The standard error measures the spread of this distribution and shows how precisely we are estimating the population mean.

Suppose we have a dataset representing expert ratings of chocolate bars, rated on a scale from 1 (unpleasant) to 5 (elite), for chocolate bars manufactured in a number of different countries.

Population vs. Sample

- **Population:** The entire group we're interested in.
    - All chocolate bars produced in each country.
- **Sample:** A smaller subset of the population that we actually measure.
    - Example: 125 chocolate bars randomly selected from the population.

We often use samples to estimate population properties because measuring the entire population is impractical.

For Figure 5.2, we have all ratings for 125 chocolate bars manufactured in Canada. Underneath each sample, which is shown as a strip chart of jittered dots, we see the sample mean plus/minus the standard deviation of the sample, the sample mean plus/minus the standard error, and 80%, 95%, and 99% confidence intervals. All five error bars are derived from the variation in the sample, and they are all mathematically related, but they have different meanings. They are also visually quite distinct.

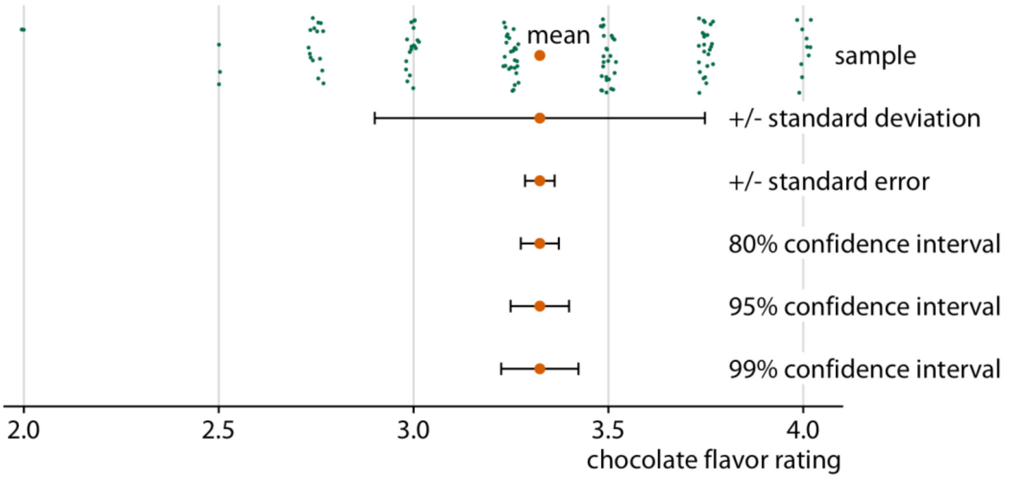

Figure 5.2 Sample, the sample mean, the standard deviation, the standard error, and different confidence intervals relate to each other using chocolate-bar ratings as an example. The jittered green dots represent expert ratings. The large orange dot shows the average rating. Below several types of error bars are represented: twice the standard deviation, twice the standard error, and the 80%, 95%, and 99% confidence intervals for the mean.

z-Distribution vs. t-Distribution

- **z-Distribution (Normal Distribution):** A bell-shaped curve used when the sample is large.
- **t-Distribution:** Similar to the z-distribution but has heavier tails. As *n* increases, the t-distribution approaches the z-distribution (see Figure 5.3).

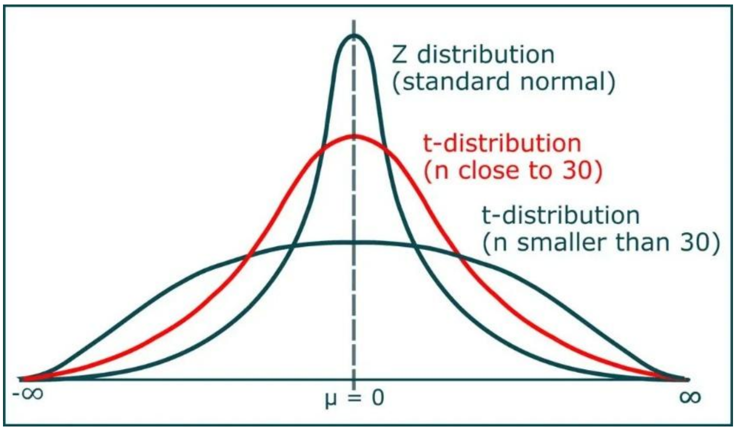

Figure 5.3 Difference between z-distribution and t-distribution.

For a dataset: $x_1, x_2, \ldots, x_n$, these measures are computed as follows:

Measure	Formula	
Sample Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	
Std. deviation	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$	
Std. error	$SE = s/\sqrt{n}$	

Mean ± SD	$\bar{x} \pm s$	- Error bar endpoints: $\bar{x} - s$ and $\bar{x} + s$ - Shows the spread of individual data points.
Mean ± SE	$\bar{x} \pm SE$	- Error bar endpoints: $\bar{x} - SE$ and $\bar{x} + SE$ - Shows uncertainty about the mean, not the data.
Confidence Interval (CI)	Large dataset: $\bar{x} \pm z_{\alpha/2}.SE$ Small dataset: $\bar{x} \pm t_{\alpha/2,df}.SE$	- α is the significance level - $z_{\alpha/2}$ is the critical value from the z-distribution - df degree of freedom (df) - $t_{\alpha/2,df}$ is the critical value from the t−distribution
80% CI	$\bar{x} \pm t_{0.10,n-1}.SE$ $\bar{x} \pm z_{0.10}.SE$	$\alpha = 0.20$
95% CI	$\bar{x} \pm t_{0.025,n-1}.SE$ $\bar{x} \pm z_{0.025}.SE$	$\alpha = 0.05$
99% CI	$\bar{x} \pm t_{0.005,n-1}.SE$ $\bar{x} \pm z_{0.005}.SE$	$\alpha = 0.01$

***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

The standard error is roughly the sample standard deviation divided by the square root of the sample size. Confidence intervals are then created by multiplying this standard error by a constant. For instance, a 95% confidence interval reaches about two standard errors above and below the mean. This means that larger samples produce smaller standard errors and narrower confidence intervals, even when the standard deviation stays the same.

We can see this when comparing chocolate-bar ratings from Canada and Switzerland (Figure 5.3). Although the two countries have almost similar mean ratings and standard deviations, Canada has 125 rated bars while Switzerland has only 38. As a result, the confidence intervals for the Swiss bars are much wider. More data → smaller standard error → narrower confidence intervals → higher certainty about the mean.

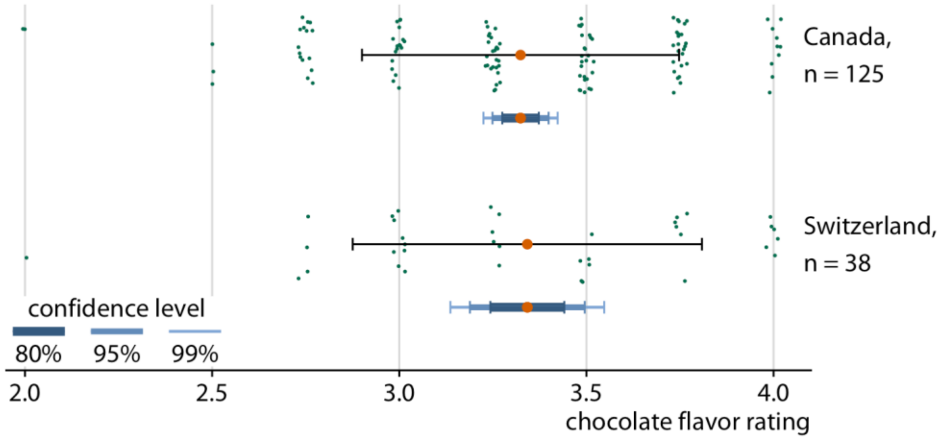

Figure 5.3 Confidence intervals become wider when the sample size is smaller. Although Canadian and Swiss chocolate bars have similar mean ratings and similar standard deviations (shown with black error bars), far fewer Swiss bars were rated. Because Canada has more than three times as many ratings, its confidence intervals are much narrower. In contrast, the mean rating for Swiss bars has much wider confidence intervals.

In Figure 5.3, three confidence intervals are shown at once, with darker colors and thicker lines used for the intervals representing lower confidence levels. This style is called graded error bars. The grading helps readers understand that there is a range of possible values, rather than a single fixed value.

Error bars are useful because they let us display many estimates along with their uncertainties at the same time. This makes them popular in scientific publications, where the goal is often to communicate a lot of information to an expert audience. For example, Figure 5.4 shows the mean chocolate ratings and corresponding confidence intervals for bars made in six different countries.

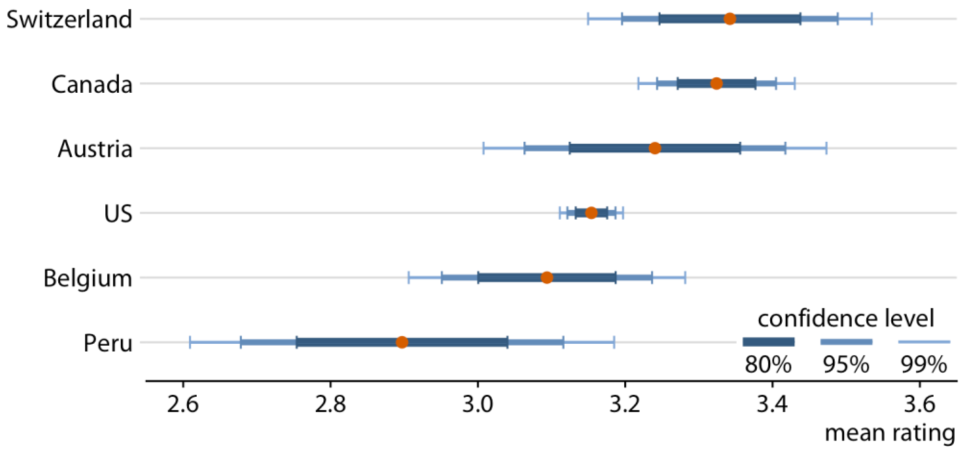

Figure 5.4 Mean chocolate flavor ratings and their confidence intervals for bars produced in six different countries.

When we look at Figure 5.4, we might ask what it tells us about differences in mean ratings. Canadian, Swiss, and Austrian bars have higher mean ratings than US bars, but are these differences statistically significant? In statistics, "significant" means that we can be confident the difference is not just due to random chance. Since only a limited number of Canadian and US bars were rated, it's possible that raters happened to evaluate more of the better Canadian bars or fewer of the better US bars, making it look like Canadian bars are better even if the true average is the same.

It's hard to judge significance from Figure 5.4 because both the Canadian and US mean ratings have uncertainty, and both uncertainties affect whether the means truly differ. The correct method is to calculate confidence intervals for the differences between means. If a confidence interval for a difference excludes zero, the difference is statistically significant at that confidence level.

For the chocolate ratings, only Canadian bars are significantly higher-rated than US bars (Figure 5.5). The 95% confidence interval for Swiss bars barely includes zero, so the observed difference from US bars could be due to chance. Austrian bars show no evidence of being rated higher than US bars.

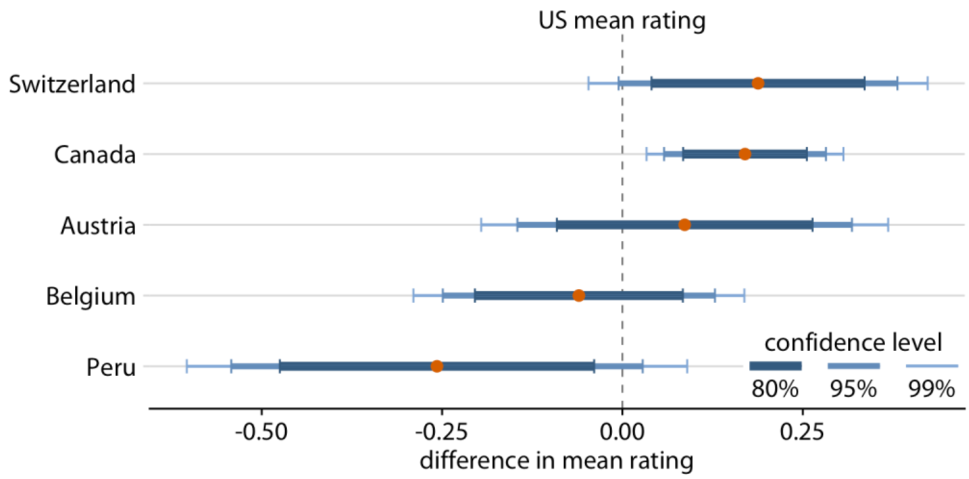

Figure 5.5 Mean chocolate ratings for manufacturers from five countries, compared to the mean rating of US bars. Only Canadian bars are rated significantly higher than US bars. For the other countries, the mean ratings are not significantly different from the US at the 95% confidence level.

Confidence levels have been adjusted for multiple comparisons using Dunnett's method. Dunnett's method is used when we have one control group (e.g., US chocolate bars) and compare multiple other treatment groups to that control (Canada vs US, Switzerland vs US, Austria vs US, etc.)
The confidence interval becomes:

$$(\bar{x}_i - \bar{x}_{control}) \pm t_{Dunnett}.SE$$

Where:

- $t_{Dunnett}$ is a *special critical value*
- It depends on the number of groups being compared
- It is larger than the usual t-value
- It makes CIs wider → reduces false discoveries

In the previous figures, we used two types of error bars: graded and simple. Many other variations are possible. For instance, error bars can be drawn with or without caps (Figures 5.6a,c vs. 5.6b,d). Each choice has pros and cons.

- **Graded error bars** show different ranges for multiple confidence levels, but they add visual complexity. In a dense or complex figure, simple error bars may be easier to read.
- **Caps** indicate the exact end of an error bar, while no caps emphasize the full range of the interval. Again, caps increase visual noise, so in figures with many error bars, omitting them can improve clarity.

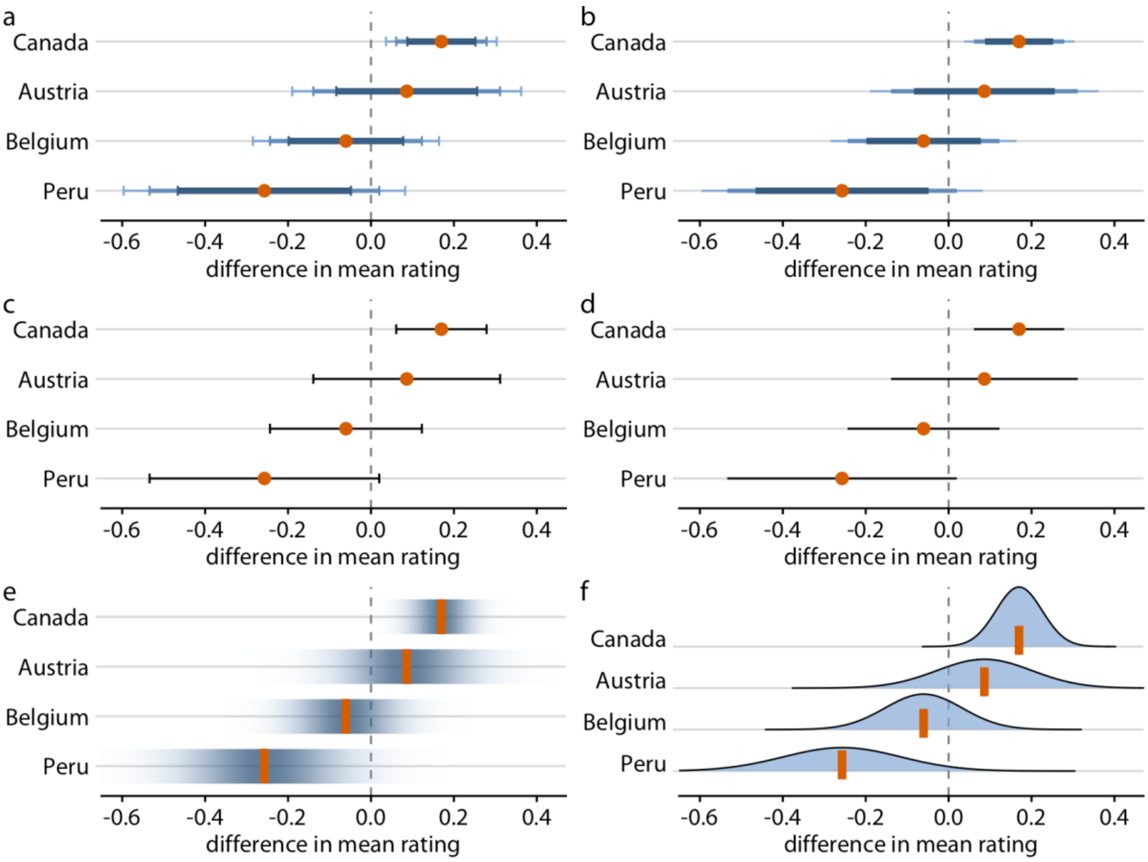

Figure 5.6 shows the mean chocolate ratings for manufacturers from four countries, compared to the mean rating of US chocolate bars. Each panel presents the same uncertainty information using a different visualization: (a) graded error bars with caps, (b) graded error bars without caps, (c) single-interval error bars with caps, (d) single-interval error bars without caps, (e) confidence strips, and (f) confidence distributions.

As an alternative to error bars, we can use **confidence strips** that gradually fade out (Figure 5.6(e)). These strips better illustrate the likelihood of different values, but they are harder to interpret because the viewer must visually integrate the shading to determine specific confidence levels. For example, Figure 5.6(e) might misleadingly suggest that the mean rating for Peruvian chocolate bars is significantly lower than that of US bars, which is not true. Similar challenges occur with **explicit confidence distributions** (Figure 5.6(f)), where it is difficult to judge exact confidence levels by looking at the area under the curve.

For simple 2D figures, error bars have a key advantage over more complex uncertainty displays: they can be easily combined with almost any type of plot. This makes it straightforward to show uncertainty alongside other visualizations. For example, we can add error bars to a bar plot to indicate variability (Figure 5.7), a common approach in scientific publications. Error bars can also be drawn in both the x and y directions on a scatterplot (Figure 5.8).

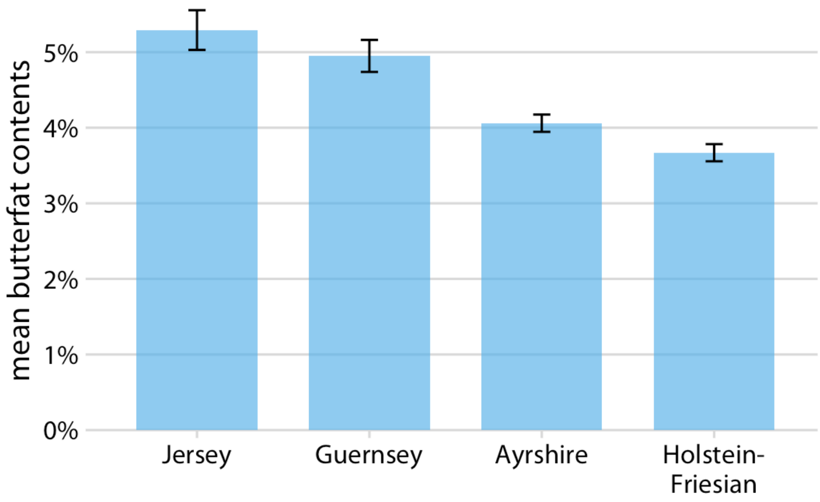

Figure 5.7 Mean butterfat content of milk from four cattle breeds. Error bars represent ± one standard error of the mean.

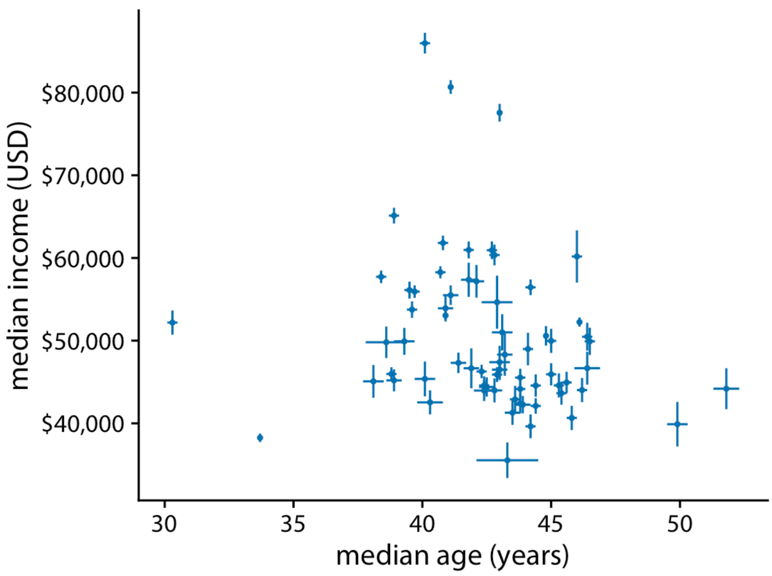

Figure 5.8 Median income versus median age for 67 counties in Pennsylvania. Error bars indicate 90% confidence intervals.

5.2. Visualizing the Uncertainty of Curve Fits

In Chapter 3, we saw how to show a trend by fitting a line or curve to data. These trend estimates are uncertain, so we often show a confidence band around the trend line (Figure 5.9). The band shows a range of fit lines that are compatible with the data.

Even a straight-line fit produces a curved confidence band because the line can vary in intercept (up or down) and slope (tilt). We can illustrate this by drawing several alternative fit lines from the distribution of possible slopes and intercepts. Figure 5.10 shows 15 such lines. Although each line is straight, the combination of different slopes and intercepts forms the curved shape of the confidence band.

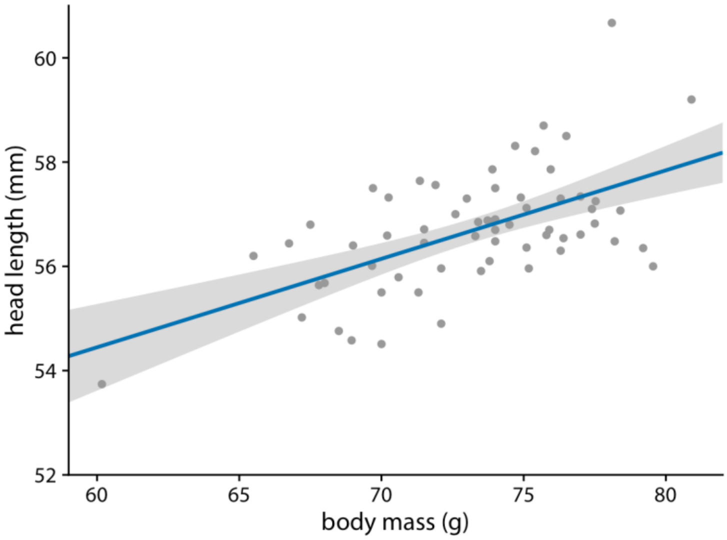

Figure 5.9 Head length vs. body mass for male blue jays. The blue line shows the best-fit line, and the gray band around it shows the 95% confidence interval, representing the uncertainty in the fit.

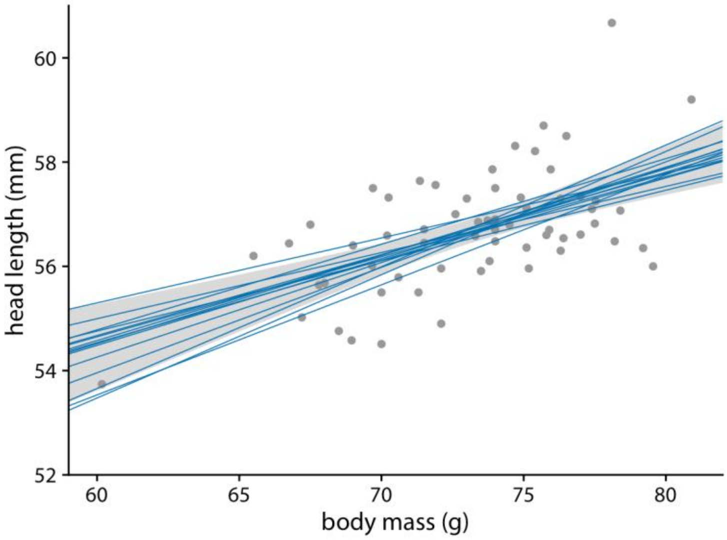

Figure 5.10 Head length vs. body mass for male blue jays. The blue lines show alternative fit lines.

- **Trend estimation:**
  - Regression models (linear or nonlinear) try to estimate the relationship between a predictor variable x and a response variable y.
  - For linear regression:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where ε represents random noise (unexplained variation).

- The predicted values $\hat{y}$ are our best estimate of the trend.

- **Residuals and Variance:**
  - Residuals = difference between observed and predicted values.

$$\text{residual}_i = y_i - \hat{y}_i$$

  - The residual variance σ^2 estimates the variability of the data around the fitted line.
  - Dividing by $n-2$ accounts for the two parameters estimated in a simple linear regression (β_0 and β_1).

$$\sigma^2 = \frac{\sum_{i=1}^{n} \text{residual}_i{}^2}{n-2}$$

- **Standard error:**

The **standard error** measures how much the predicted value $\hat{y}$ could vary if we repeated the experiment.

$$SE_{\hat{y}} = \sqrt{var(\hat{y})} = \sqrt{\sigma^2(1/n + (x-\bar{x})^2 / \sum(x_i - \bar{x})^2)}$$

$(x-\bar{x})^2/\sum(x_i-\bar{x})^2$accounts for how far the x-value is from the mean predictions are more uncertain at the edges of the data.

- **Confidence interval:** Uses the t-distribution to reflect uncertainty from finite data.

$$\hat{y} \pm t_{\alpha/2,n-2} \cdot SE$$

To create a confidence band, we first choose a confidence level. Similar to error bars, it can be helpful to show multiple confidence levels simultaneously. This gives rise to a graded confidence band (Figure 5.11), which displays several levels of confidence at once. A graded band makes the uncertainty more tangible and encourages the reader to recognize that the data may support a range of alternative trend lines.

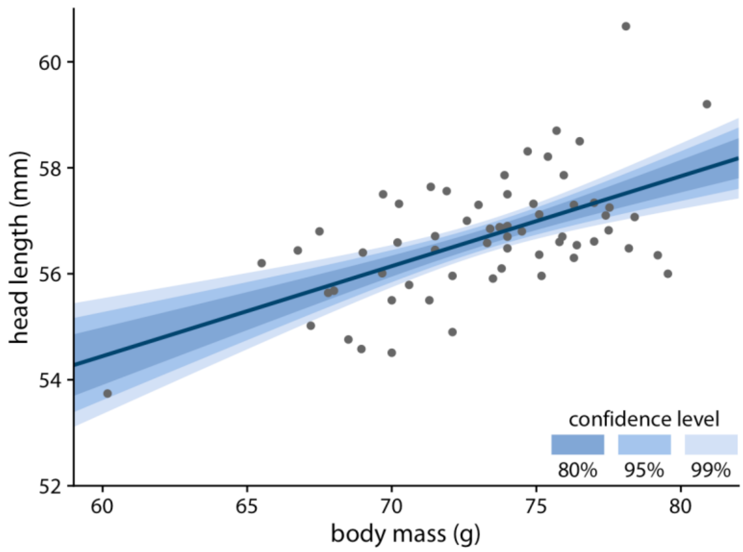

Figure 5.11 Head length vs. body mass for male blue jays. Graded confidence bands show the uncertainty in the trend estimate, similar to error bars.

Confidence bands can also be drawn for nonlinear curve fits. While they look appealing, they can be harder to interpret (Figure 5.12). In Figure 5.12(a), it might seem that the band simply shows the blue curve shifting up and down or slightly deforming. However, Figure 5.12(b) shows that the band actually represents a family of curves that are often much wigglier than the best-fit curve in (a). In general, for nonlinear fits, uncertainty involves not just vertical shifts but also variations in the curve's shape.

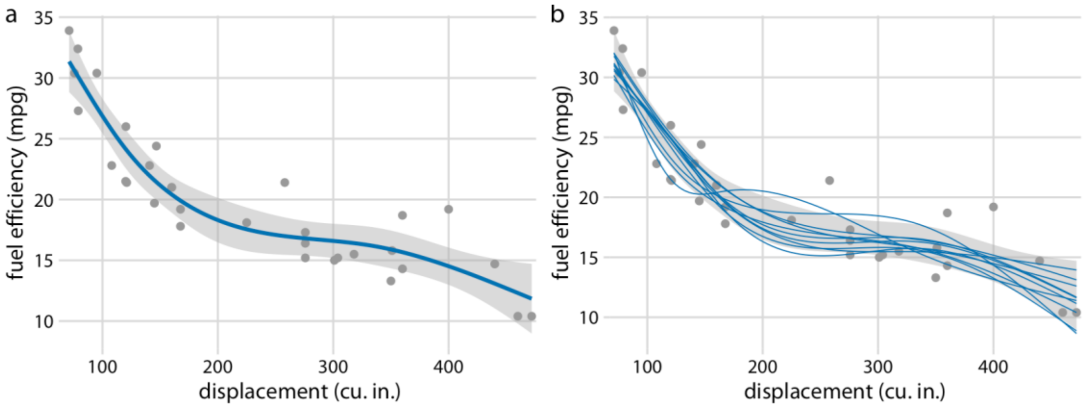

Figure 5.12 Fuel efficiency vs. displacement for 32 cars (1973–74 models). Each dot is a car. The smooth lines are cubic regression splines with 5 knots. (a) Best-fit spline with confidence band (b) Alternative fits

5.3. Visualizing the Uncertainty of outcomes

Showing individual possible outcomes, is called a *discrete outcome visualization*. Presenting probabilities as frequencies is known as *frequency framing*, it expresses uncertainty in terms of simple, understandable counts of outcomes.

For example, by drawing squares in different colors. In Figure 5.13, this method is used to show three probabilities: a 1% chance of success, a 10% chance, and a 40% chance.

To interpret the figure, imagine you must pick a dark square without seeing the grid in advance, like choosing with your eyes closed. You can quickly sense that picking the single dark square in the 1% case is very unlikely. It's still unlikely in the 10% case, but in the 40% case your chances look much better.

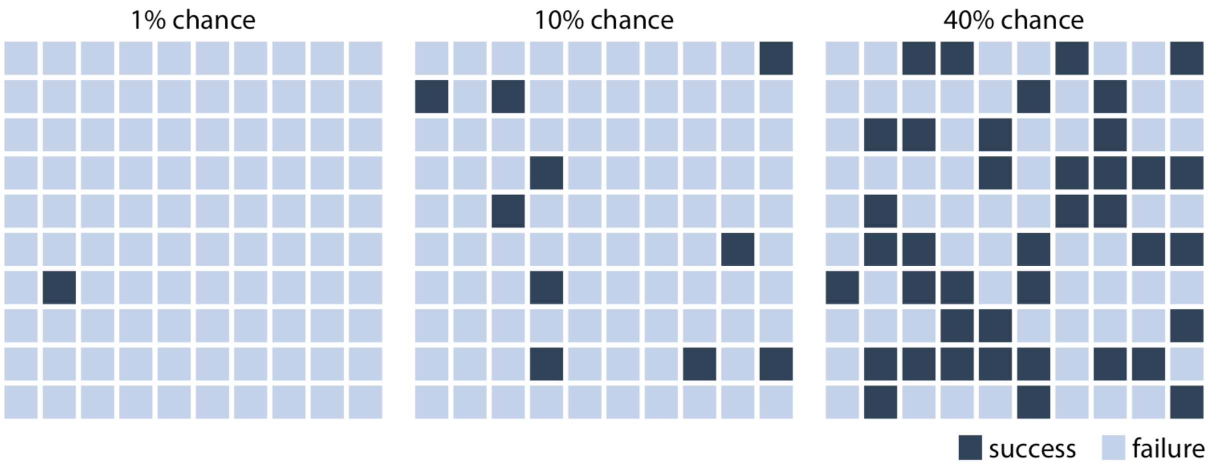

Figure 5.13 Uncertainty of outcomes. Each 100-square grid represents outcomes of a random trial, with dark squares indicating success. The 1%, 10%, and 40% success cases contain 1, 10, and 40 dark squares, respectively. Randomly distributing the dark squares creates a visual sense of uncertainty in any single trial.

If we are only interested in two discrete outcomes, success or failure, then a visualization such as Figure 5.13 works fine. However, often we are dealing with more complex scenarios where the outcome of a random trial is a numeric variable. One common scenario is that of election predictions, where we are interested not only in who will win but also by how much. Let's consider a hypothetical example of an upcoming election with two parties, the yellow party and the blue party. Assume we hear on the radio that the blue party is predicted to have a 1 percentage point advantage over the yellow party, with a margin of error of 1.76 percentage points. What does this information tell you about the likely outcome of the election? It is human nature to hear "the blue party will win," but reality is more complicated. First, and most importantly, there are a range of different possible outcomes. The blue party could end up winning with a lead of two percentage points, or the yellow party could end up winning with a lead of half a percentage point. The range of possible outcomes with their associated likelihoods is called a *probability distribution*, and we can draw it as a smooth curve that rises and then falls over the range of possible outcomes (Figure 5.14). The higher the curve for a specific outcome, the more likely that outcome is. Probability distributions are closely related to the histograms and kernel densities.

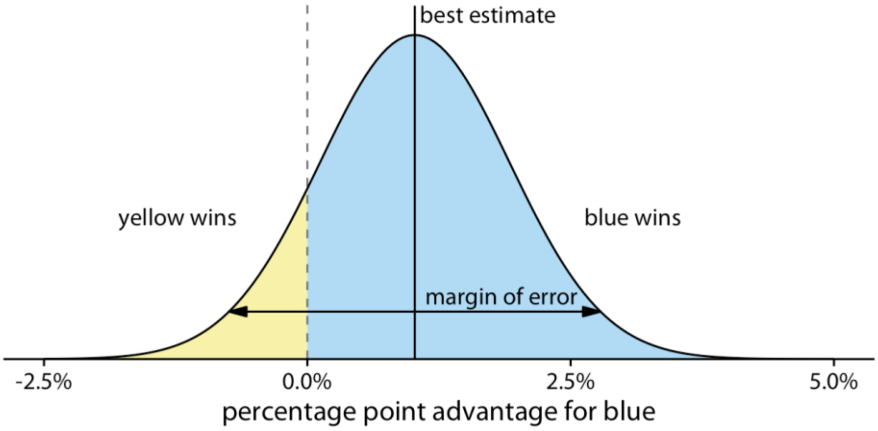

Figure 5.14 shows a hypothetical election forecast. The blue party is expected to lead by 1 percentage point, with a 95% margin of error of ±1.76 points. The blue-shaded area (87.1%) shows outcomes where blue wins, and the yellow-shaded area (12.9%) shows outcomes where yellow wins. This means blue has an 87% chance of winning.

We need to define the mean μ and the standard deviation σ in order to plot the density plot. Here, the mean represents the expected difference, which is the 1 percentage point advantage of the blue party over the yellow party. The margin of error (MOE) is tied to the confidence interval, which indicates that with a level of confidence of 95%, the true value of the difference lies within:

$$\text{Confidence interval (CI)} = (\mu - MOE, \mu + MOE) = (1\text{-}1.76,\ 1\text{+}1.76) = (\text{-}0.76,\ 2.76)$$

The standard deviation σ is estimated as: $MOE = z.\sigma$, where z is the z-score at the confidence level of 95%, which is equal to 1.96. So, $\sigma = \frac{1.76}{1.96} \approx 0.90$

By doing some math, we can calculate that for our made-up example, the chance of the yellow party winning is 12.9%. So, the chance of yellow winning is a tad better than the 10% chance scenario shown in Figure 5.13. If you favor the blue party, you may not be overly worried, but the yellow party has enough of a chance of winning that it might just be successful. If you compare Figure 5.14 to Figure 5.13, you may find that Figure 5.13 creates a much better sense of the uncertainty in outcome, even though the shaded areas in Figure 5.14 accurately represent the probabilities of blue or yellow winning. This is the power of a discrete outcome visualization. Research in human perception shows that we are much better at perceiving, counting, and judging the relative frequencies of discrete objects, as long as their total number is not too large, than we are at judging the relative sizes of different areas.

We can combine the discrete outcome nature of Figure 5.13 with a continuous distribution as in Figure 5.14 by drawing a *quantile dot plot*. In the quantile dot plot, we subdivide the total area under the curve into evenly sized units and draw each unit as a circle. We then stack the circles such that their arrangement approximately represents the original distribution curve (Figure 5.15).

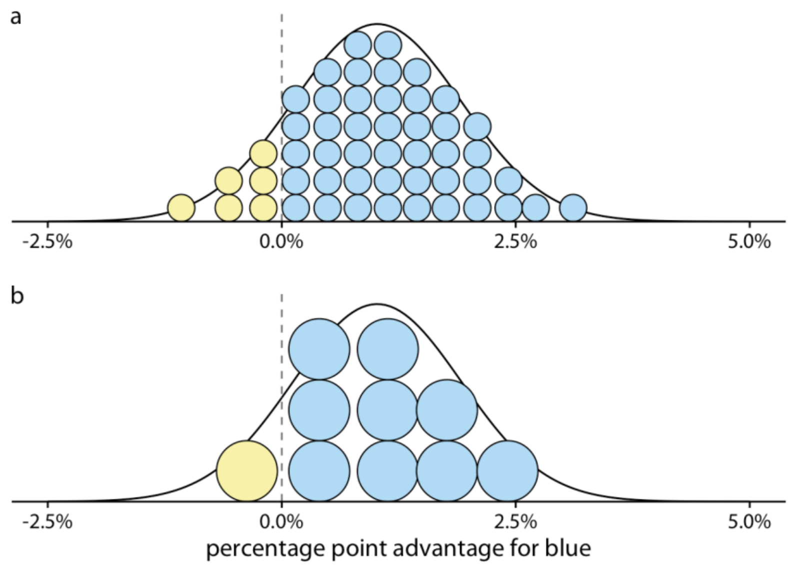

Figure 5.15 Quantile dot plot representations of the election outcome distribution of Figure 5.14. (a) The smooth distribution is approximated with 50 dots representing a 2% chance each. The 6 yellow dots thus correspond to a 12% chance, reasonably close to the true value of 12.9%. (b) The smooth distribution is approximated with 10 dots representing a 10% chance each. The 1 yellow dot thus corresponds to a 10% chance, still close to the true value.

As a rule, quantile dot plots work best with a small to moderate number of dots. Too many dots make them appear as a continuous mass rather than discrete units, reducing the benefit of using a discrete plot. Figure 8.3 compares 50 dots (Figure 5.15(a)) and 10 dots (Figure 5.15(b)). The 50-dot version better represents the true probability distribution, but it's hard to see individual dots. The 10-dot version more clearly shows the relative chances of blue or yellow winning.

While the 10-dot plot slightly underrepresents yellow's chance by 2.9 percentage points, sacrificing a bit of precision can improve human perception, especially for a general audience. A mathematically exact plot that isn't easily understood is often less useful in practice.

5.4. Hypothetical Outcome Plots

When uncertainty is shown in a static plot, people might mistakenly think that some part of the display is a fixed property of the data. To avoid this, we can animate the uncertainty by showing many different but equally likely versions of the plot. This approach is called a hypothetical outcome plot (HOP). HOPs can't be used in print, but they work very well online as GIFs or videos, and they're also useful during oral presentations.

To show how a HOP works, let's revisit the chocolate bar ratings example. When you're choosing chocolate at the store, you probably don't care about mean ratings or their uncertainty. Instead, you might wonder: if I randomly choose one Canadian bar and one US bar, which is more likely to taste better?

To answer this, we could repeatedly pick one Canadian and one US bar at random from the dataset, compare their ratings, and record which one comes out on top. If we repeat this many times, we find that the Canadian bar wins about 53% of the time, while the US bar wins or ties about 47% of the time. A HOP visualizes this idea by cycling through several of these random draws and showing the ranking of the two bars in each one (Figure 5.16).

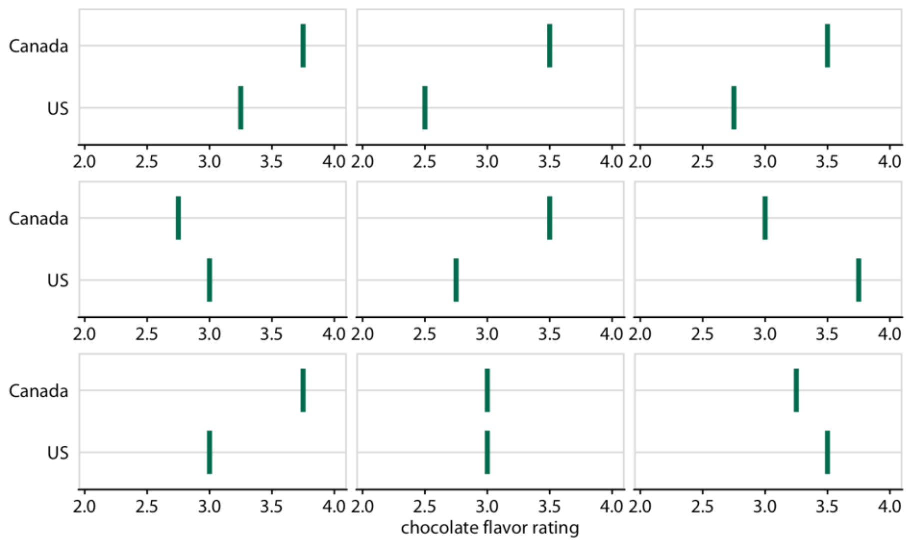

Figure 5.16 Hypothetical outcome plot for the ratings of Canadian and US chocolate bars. Each vertical green bar shows the rating of one bar, and each panel compares two randomly selected bars, one from Canada and one from the US. In a real HOP, these panels would appear one after another as an animation rather than being shown side by side.

As another example, look at the variation among the equally likely trend lines. Because all the lines are drawn on top of each other, we mostly see the general region they occupy, much like a confidence band, and it becomes hard to distinguish any single trend line. Turning this figure into a HOP lets us show one trend line at a time (Figure 5.17).

When creating a HOP, you might wonder whether to switch sharply between outcomes (like flipping slides) or to smoothly animate from one to the next (for example, gradually morphing one trend line into another). Research on this is still ongoing, but some studies suggest that smooth transitions can make it harder for viewers to understand the underlying probabilities. If we use smooth animations, it may help to keep them very fast or use fade-in/fade-out effects instead of gradual shape changes.

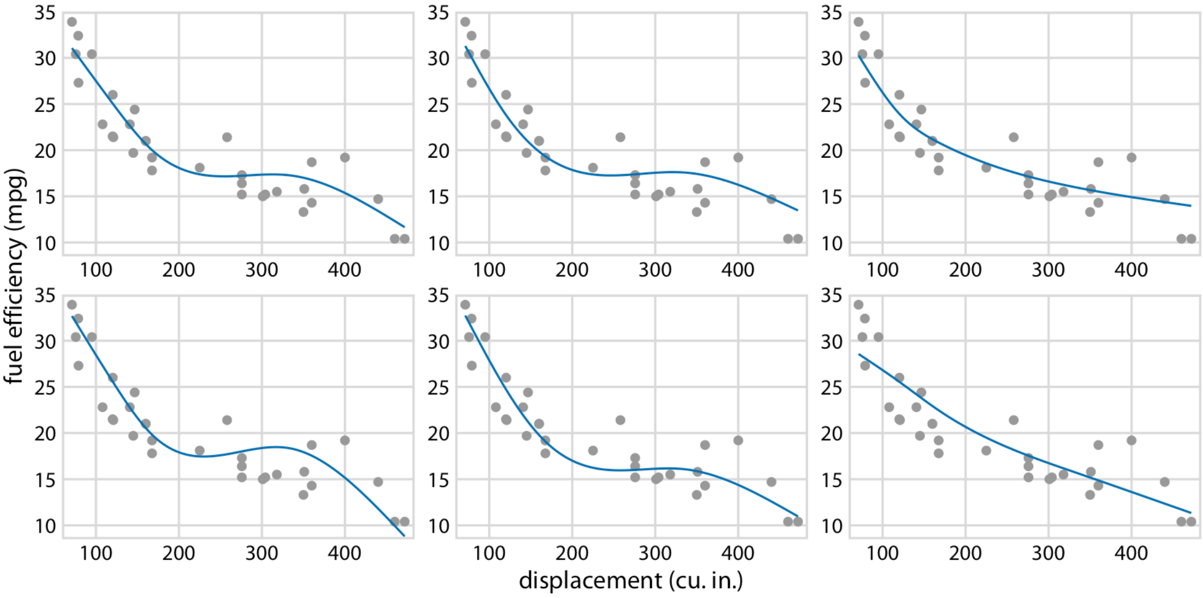

Figure 5.17 Hypothetical outcome plot for fuel efficiency versus engine displacement. Each dot corresponds to a single car, and the smooth curves come from fitting a cubic regression spline with five knots. Each curve in each panel represents one possible fit, sampled from the posterior distribution of the model's parameters. In a real HOP, these panels would appear sequentially as an animation rather than side by side.

One important point when creating a HOP is ensuring that the displayed outcomes accurately reflect the true distribution of possibilities. Otherwise, the visualization can be misleading. For instance, in the chocolate ratings example, if we randomly draw only 10 pairs and 7 of them show the US bar scoring higher, the HOP would wrongly suggest that US bars are usually better. We can avoid this by using a large number of samples—reducing the chance of bias—or by checking that the displayed outcomes are representative. For Figure 5.16, for example, we confirmed that the proportion of times the Canadian bar won was close to the true rate of 53%.