Variability of Samples

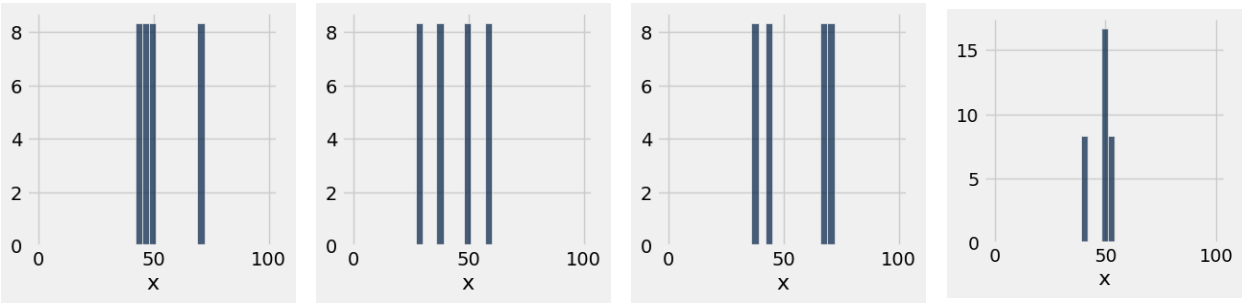
Note 02

Data C8: Foundations of Data Science | UC Berkeley, Fall 2025

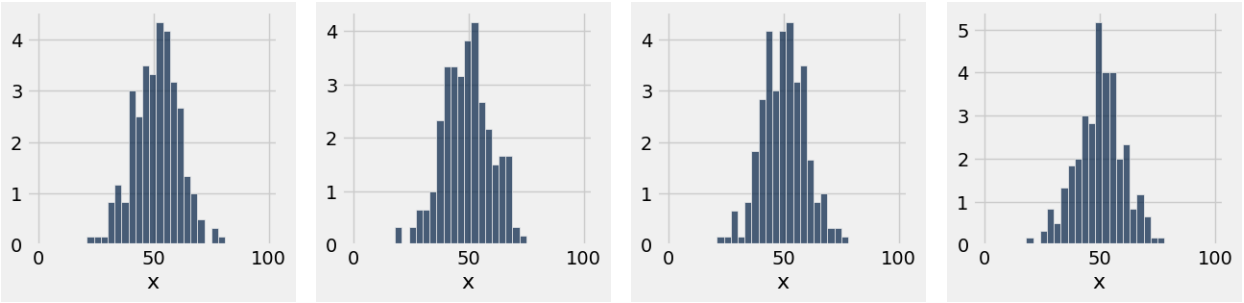
Since sampling from a population is a random procedure, we will inevitably obtain samples that differ from each other. We will discuss this variability in terms of the sample mean and quantify it by computing the *standard deviation of sample means*.

Let's first think about this intuitively. If we were to take small samples from the population, there is a higher chance that we may get more skewed samples. For example, a sample size of $n = 4$ has a higher chance that its values may be very small or large. This means that the sample means could vary significantly. On the other hand, if we took very large samples from the population, then the samples would be much more representative of the underlying distribution, and for the sample to be skewed is much more unlikely. In the example graphs below, the underlying distribution was a normal distribution centered at $x = 50$.

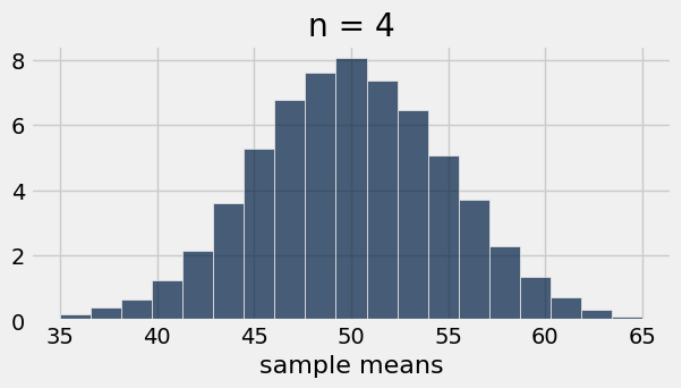
$n = 4$

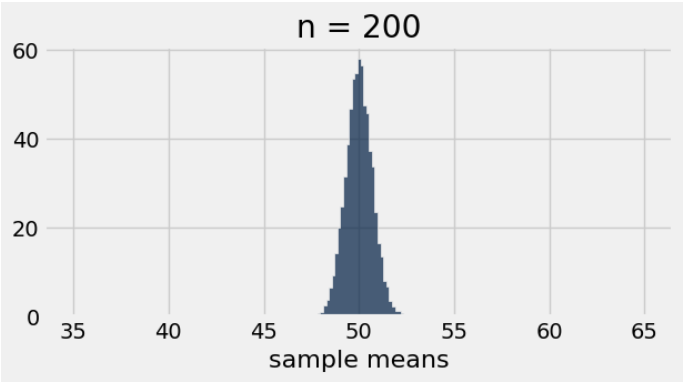


$n = 200$



Notice how the small sample size ($n = 4$) gives very different samples, and therefore the sample means are more spread out. The larger samples look more representative of the original distribution (a normal distribution), and have sample means that are less spread out. Thus, taking *small* samples leads to *high* variability in the sample mean, and vice versa – which is shown empirically below, and more concretely defined by the relationship between the *standard deviation of sample means* and the *sample size*.





Formula

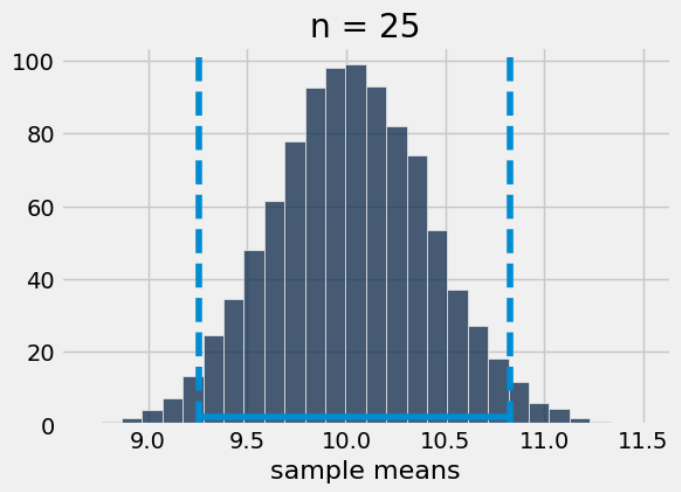
$$\text{SD of sample means} = \frac{\text{SD of population}}{\sqrt{\text{sample size}}}$$

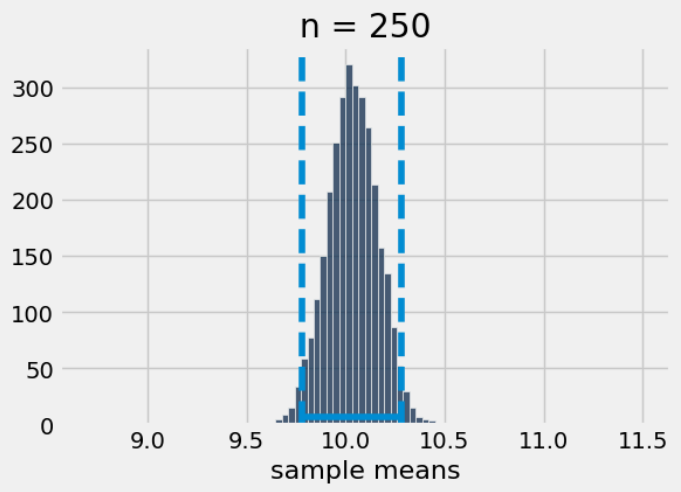
Note that this is the standard deviation of sample means across *all* possible samples, **not** of a single sample.

Choosing Sample Sizes

In practice, we often need to think *backwards*. We start with some desired *accuracy* and ask "How large should our samples be so that a 95% confidence interval is about this wide?" We can use sample size to *control* the accuracy of our sample means: larger samples give more accurate estimates, which show up as narrower confidence intervals.

Recall that a 95% confidence interval is a range in which we can expect the parameter to be in, 95% of the time. The confidence therefore models the variability in our sample: lower variability means higher accuracy, which corresponds to a smaller interval width.





But how do we describe accuracy in terms of the interval width? Recall from the Central Limit Theorem that the distribution of sample means will be roughly normal. This means that we can describe a confidence interval as the middle 95% of a normal distribution, and instead of using percentiles to calculate the interval, we already know that the mean $\pm$ 2 SDs contains roughly 95% of data.

Therefore, taking 2 SDs on either side, the width of a 95% confidence interval is roughly

$$\text{width} \approx 4 \cdot \text{SD of sample means} = 4 \cdot \frac{\text{SD of population}}{\sqrt{\text{sample size}}}$$

If we have a desired accuracy in mind, we can calculate the width and determine the sample size necessary to obtain this accuracy.

> Example
> **Survey Planning**
> Suppose we were planning a survey for the time people spend in a grocery store, with a population standard deviation of 10 minutes. We want a 95% confidence interval for the mean time with a maximum width of 4 minutes.
>
> Using the formula above, we have that
>
> $$\text{width} = 4 \cdot \frac{\text{SD of population}}{\sqrt{\text{sample size}}} \leq 4 \implies \text{sample size} \geq \left(\frac{4 \cdot \text{SD of population}}{4}\right)^2 = 10^2 = \boxed{100}$$
>
> This means that we need a sample size greater than 100 to give a 95% interval less than 4 minutes wide.