Variability of Sampling Distributions

An extremely valuable property to know of sampling distributions is knowing how variable sample properties are from sample to sample. And doing so can give us some comfort and confidence in the particular sample we get.
For example, if we expect samples to differ a lot from one another (under the hypothetical scenario where we sample multiple times), we shouldn't put too much confidence in any particular sample we get. But if we expect samples to be quite similar to one another, we could be quite confident in the inferences we make from a particular sample.

Standard Errors - Standard Deviations of Sampling Distributions

We have a way to measure how different sample properties might be. The measurement we use is the same as we use to measure variability of data or variability of populations. We can use the standard deviation.

A standard deviation is a measure of how different individuals are from the average. Sometimes the standard deviation is described as a measure of "spread". Standard deviations are *approximately* the expected distance of an individual from the average. To see why, let's break down the components of the standard deviation.

Imagine we want to measure how different individuals are from each other. You could measure how different each individual is from all others, but that would be quite unwieldy (for *n* individuals, we would have *n*(*n*-1)/2 unique differences between individuals, which gets large pretty quickly). Instead, let's measure how different every individual is from the average. Let X_i be the character for individual *i* in the population and let $\bar{X}$ be the population mean.

The average is defined as

$$\bar{X} = \frac{1}{n}\sum_{i}^{n} X_i$$

The difference between an individual character and the average is

$$x_i - \bar{x} \leftarrow \text{the difference between individual } i\text{'s character and the average.}$$

Let's just take the average of the differences. That should get us how different individuals are from each other on average, right? Taking the average of the differences is

$$\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\text{definition of } \bar{x}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} \bar{x}}_{\text{adding up } \bar{x} \text{ n times then dividing by n.}}$$

↗ average differences

$$= \frac{1}{n}(n \cdot \bar{x})$$
$$= \bar{x}$$

Thus, we have

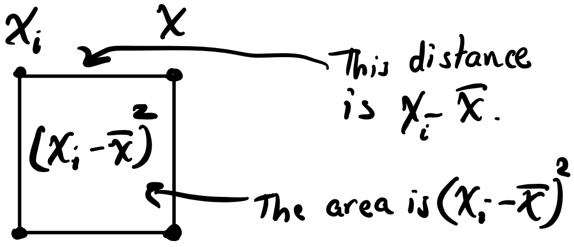
$$\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x} = \bar{x} - \bar{x} = 0$$

The problem with using the average difference is that all the positive and negative differences from the mean cancel out. We can solve this problem by using a measure of distance from the mean that doesn't depend on whether a value is above or below the mean. One such kind of distance is found by taking the square of the distance. That is, the squared distance of an individual from the mean is

Squared Distance From the Mean.

$$\underbrace{(x_i - \bar{x})}_{\text{Distance From Mean}}^{2} \leftarrow \text{squared.}$$

Graphically, this makes the distance an area.


x_i
$\bar{x}$
This distance is $x_i - \bar{x}$.
$(x_i - \bar{x})^2$
The area is $(x_i - \bar{x})^2$

Imagine now that we have many individuals. We could measure their squared distances and get an average to estimate the average squared difference among all the individuals. This is what a variance does. Let's see an example with 3 observations: 10,11, and 12.

Individual Observations

$x_1 = 10$

$x_2 = 11$

$x_3 = 12$

Mean

$$\frac{x_1 + x_2 + x_3}{3} = \frac{10+11+12}{3} = \frac{33}{3} = 11.$$

Differences From Mean

$x_1 - \bar{x} = 10 - 11 = -1$

$x_2 - \bar{x} = 11 - 11 = 0$

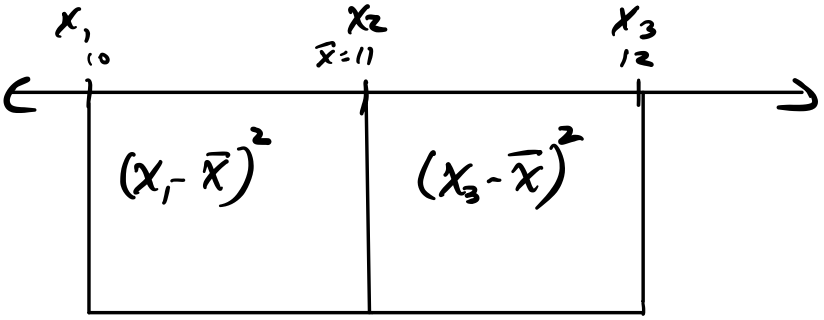
$x_3 - \bar{x} = 12 - 11 = 1$

Squared Differences From Mean

$(x_1 - \bar{x})^2 = (-1)^2 = 1$

$(x_2 - \bar{x})^2 = 0^2 = 0$

$(x_3 - \bar{x})^2 = 1^2 = 1$



Average the $(x_i - \bar{x})^2$

$$\frac{1+0+1}{3} = \frac{2}{3}$$

$\boxed{2/3}$ ← Length of one side is $\sqrt{2/3} \approx 0.82$

Length of one side is $\sqrt{2/3}$

Once we have the average of the squared deviations, we can put it back in the same units by taking the square root. This takes our unit of distance from an area back to a length. In doing so, we have calculated a **standard deviation.** For the data 10, 11, and 12, the average is 11 and the standard deviation is 0.82, which is approximately the difference we expect to see from a randomly selected individual from the average. 1/3 of the time we see 0 (because we chose 11 and that is the mean). And 2/3 of the time we see 1 (because 10 and 12 are both 1 unit away from 11).

While a standard deviation is not ***exactly*** the expected difference of individuals in the population, it is very close and is exactly the essence of what the standard deviation calculates, if not exactly the value that it calculates. The most valuable thing to realize is that standard deviations are interpretable as a measure of how different observations will be in the same units as the observations themselves.

The observations may apply to datasets or to probability distributions. With datasets, we ask how much the individuals in a sample are from one another. With probability, we can know how different we expect outcomes generated from that probability distribution will be. What this means is that **datasets and probability distributions have variances and standard deviations.** They both characterize the same thing, but they are both measured in slightly different ways.

Probability distributions are defined by a set of mutually exclusive outcomes and the probabilities associated with those outcomes. We can define all outcomes for a random variable *X* as a set. One possible value in that set is *xi*. For example, in the binomial distribution, there *x* can be any value 0,1,2,…,*n*. Thus, *xi* is the number of successes in a single draw from the binomial. We can also let the probability of *xi* be $f_X(x_i)$

The subscript *X* refers to the random variable and the lowercase *x* in the parentheses represents any particular outcome from the random variable *X*. With this notation, the mean for a probability distribution *X* is

$$\mu = \sum_i x_i f_X(x_i)$$

And the variance of the probability distribution is

$$\sigma^2 = \sum_i f_X(x_i)(x_i - \mu)^2$$

The standard deviation is just the square root of the variance, meaning the standard deviation of the probability distribution *X* is

$$\sigma = \sqrt{\sum_i f_X(x_i)(x_i - \mu)^2}$$

(Note that if we have a continuous probability distribution, we use integrals in place of the sums, which is just the way to add continuous functions. All else is the same.)

These equations for means, variances, and standard deviations, work just like their data analogues. The mean, variance, and standard deviation for a sample of data are

$$\bar{X} = \frac{1}{n}\sum_i^n X_i \quad \longleftarrow \text{Mean of data}$$

$$s^2 = \frac{1}{n-1}\sum_i^n (X_i - \bar{X})^2 \quad \swarrow$$

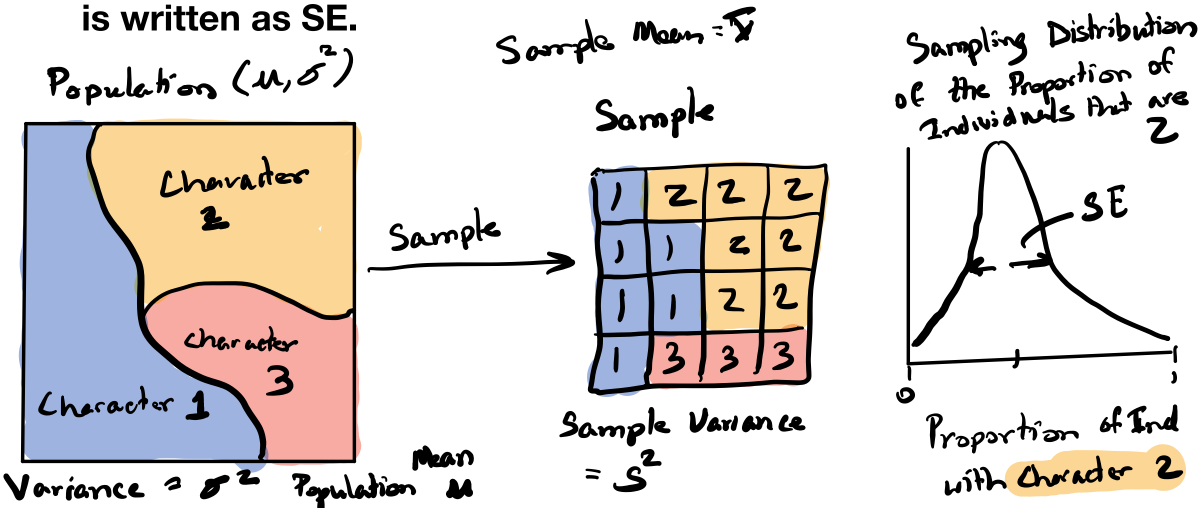
Variance of data
"Sample Variance"

$$s = \sqrt{\frac{1}{n-1}\sum_i^n (X_i - \bar{X})^2} \quad \leftarrow \text{Standard Deviation of data}$$

Understanding Variability In Different Parts of the Statistical Workflow

Statistics is all about variability, and so measuring variability turns out to be very important. And because measuring variability is important, the standard deviation (or variance) shows up in many places. Let's return to the basic framework for doing statistics.

1. We have a question and are studying a population.
2. The population has some properties we are interested in related to the question.
3. We characterize the population using a probability distribution for the individual. This probability distribution has some variance related to how different individuals in the population are. **This variability is encapsulated by the variance of the probability distribution of the individual, and is denoted by** σ^2
4. We sample the population, and measure its properties. Individuals vary within the sample similar to, but not exactly like, the population variation. **This variability is encapsulated by the sample variance, and is denoted by** s^2
5. We build a sampling distribution of the sample property we measure. This probability distribution has some variability representing how different sample estimates would be under repeated sampling. **This variability is encapsulated by the variance of the sampling distribution. The square root of the variance of the sampling distribution (i.e., its standard deviation) is called the <u>standard error of the estimate</u>, and is written as SE.**



It’s crucial to be able to distinguish these types of variability and what they describe. Here are some tips.

1. We use the sample standard deviation, *s*, to describe how different individuals are in our sample. Moreover, *s* is an estimate of how different individuals are in the population.
2. We use the standard error to describe how different estimates may be under repeated sampling of a population. For example, if we wanted to estimate the mean of a population, we measure the mean of a sample. And the standard error of the mean tells us how different we expect sample means to be for different samples.

We often visualize both sample standard deviations and standard errors with “error bars.” But they mean quite different things, and it’s important to get them straight.

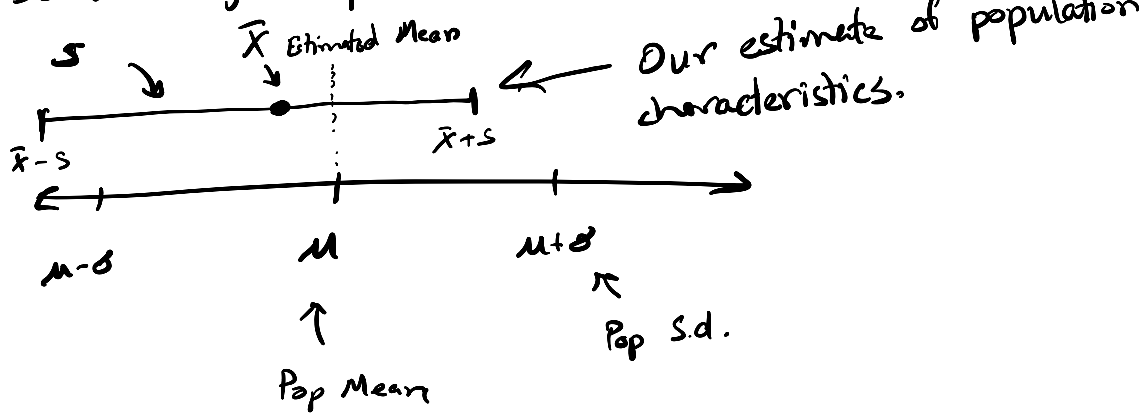
Assume we want to estimate a population mean, μ.

This population has a standard deviation σ.

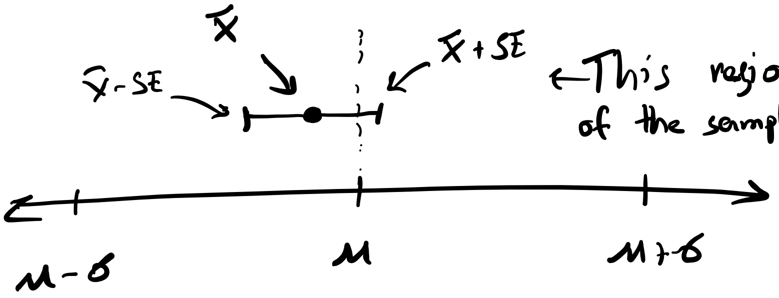
Let's collect a sample and estimate the mean $\bar{X}$ and the standard deviation with s.

Assuming random sampling s and $\bar{X}$ will be in the neighborhood of σ and μ.

So we might represent this as follows.



But the standard error of the mean is always smaller than the sample standard deviation. So it looks like this



←This region shows the features of the sampling distribution of the mean.

We could do the same thing for any other property of a population. For example, say we wanted to estimate the median of a population. We could then take a sample, estimate the median from the sample, and then construct a sampling distribution of the sample median. The variation of the sampling distribution of the median is measured by the standard error of the median. We could do the same for population variance, population skew, population kurtosis, or population minimum or maximum! Each of these things are estimated with a property of a sample and that property has a sampling distribution. **An absolutely essential component of that sampling distribution is how variable different estimates are across samples.**

If estimates vary a lot, we have to come to the conclusion that we should be unsure that are particular estimate is correct (because we are likely to get a very different answer by doing the whole procedure again!).

But if our estimate is pretty constant across samples, then we can have much more confidence that the estimate from our one sample is correct. This is because we are likely to estimate a similar value if we went back out and sampled again.

Standard Errors of the Mean

All estimates have a sampling distribution and those sampling distributions all have standard deviations that we called “standard errors”. One that we use all the time is the standard error of the mean, typically written as $SE_{\bar{X}}$

Luckily, we have a direct way to calculate the standard error of the mean. The equation is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

← Sample standard deviation

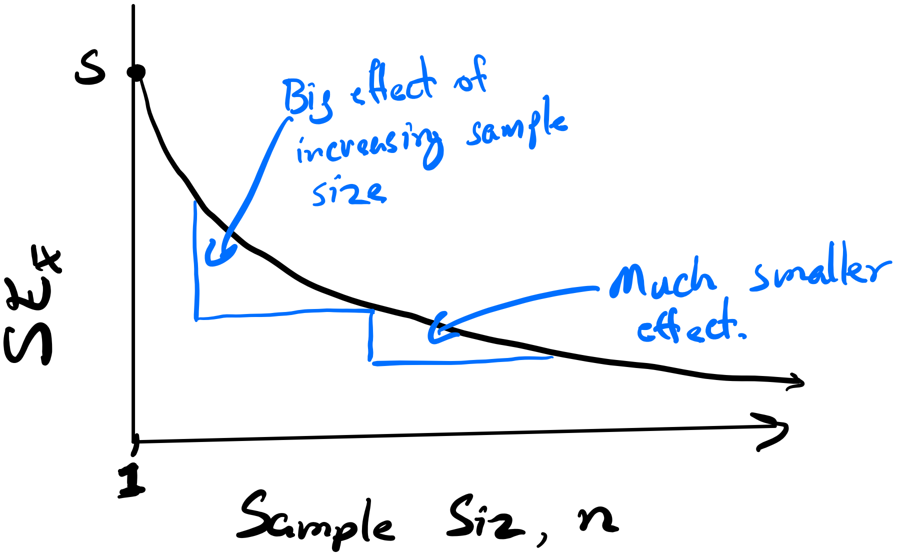
← Square root of the sample size

This equation says that the variability in sample means is higher when

1. The population is more variable (i.e., *s* is large)
2. The sample size is low (i.e., *n* is small).

This gives us a prescription to try and winnow down our estimate of the mean. We could try to reduce the sample standard deviation, but this is a property of the population and is completely out of a biologist's control. All we can do is increase the sample size, *n*, to make the standard error smaller.

Note however that increasing sample size doesn't always have the same impact. Increasing the sample size from 10 to 20 has a big effect while increasing the sample size from 100 to 110 has very little effect. The relationship between the standard error and the sample size looks like this



This also tells us that we don't need very large samples to have high confidence in our estimate if the population is relatively uniform (i.e., *s* small). We need very large samples to be confident in our estimate of the mean if the population is highly heterogeneous (i.e., *s* is large).