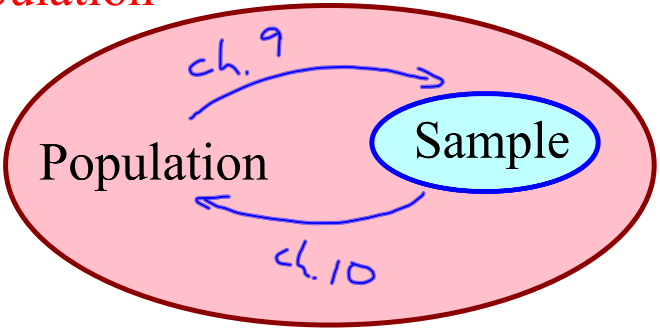
9.1 Sampling distributions

One day some papers catch fire in a wastebasket in the Dean's office. Luckily, a physicist, a chemist, and a statistician happen to be nearby. Naturally, they rush in to help. The physicist whips out a notebook and starts to work on how much energy would have to be removed from the fire in order to stop the combustion. The chemist works on determining which reagent would have to be added to the fire to prevent oxidation. While they are doing this, the statistician is setting fires to all the other wastebaskets in the adjacent offices. "What are you doing?" the Dean demands. To which the statistician replies, "To solve a problem of this magnitude, you need a large sample size."

Recall these terms:

parameter: a fixed, often unknown #
describing a population



statistic: a number known for a particular sample that describes that sample, but the value likely changes from one sample to another sample

We estimate parameters 2 ways:
point estimates (ch. 9) and
interval estimates (ch. 10).

These statistics are the best point estimates for these parameters:

	mean	variance	standard deviation	proportion
statistics	$\bar{x}$	s^2	s	$\hat{p}$
parameters	μ	σ^2	σ	p
				π

Common Sense Thing #1 Random samples are good.
A random sample should represent the population well, so a statistic from a random sample should provide a reasonable estimate of a population parameter.

Common Sense Thing #2 Statistics have error.
All statistics have some error in estimating population parameters.

Common Sense Thing #3 Statistics have distributions.
If repeated samples are taken from a population and some statistic is calculated from each sample, the statistics will vary. That is, the statistic will have a distribution with a mean and standard deviation.

Common Sense Thing #4 Larger sample size is better: less error.
A larger sample provides more information than a smaller sample so a statistic from a large sample should provide have less error than a statistic from a small sample.

Fact: Different samples yield different values of statistics. That's called **sampling variability**.

The **sampling distribution of a statistic** is

- the probability distribution of the statistic.
- the distribution of values taken by the statistic in all possible samples of the same size from the same population.

population

SRS of size n | SRS of size n

statistic based on this SRS $\bar{x}$ (or $\hat{p}$) | statistic based on this SRS $\bar{x}$ (or $\hat{p}$)

These statistics may vary from one another.

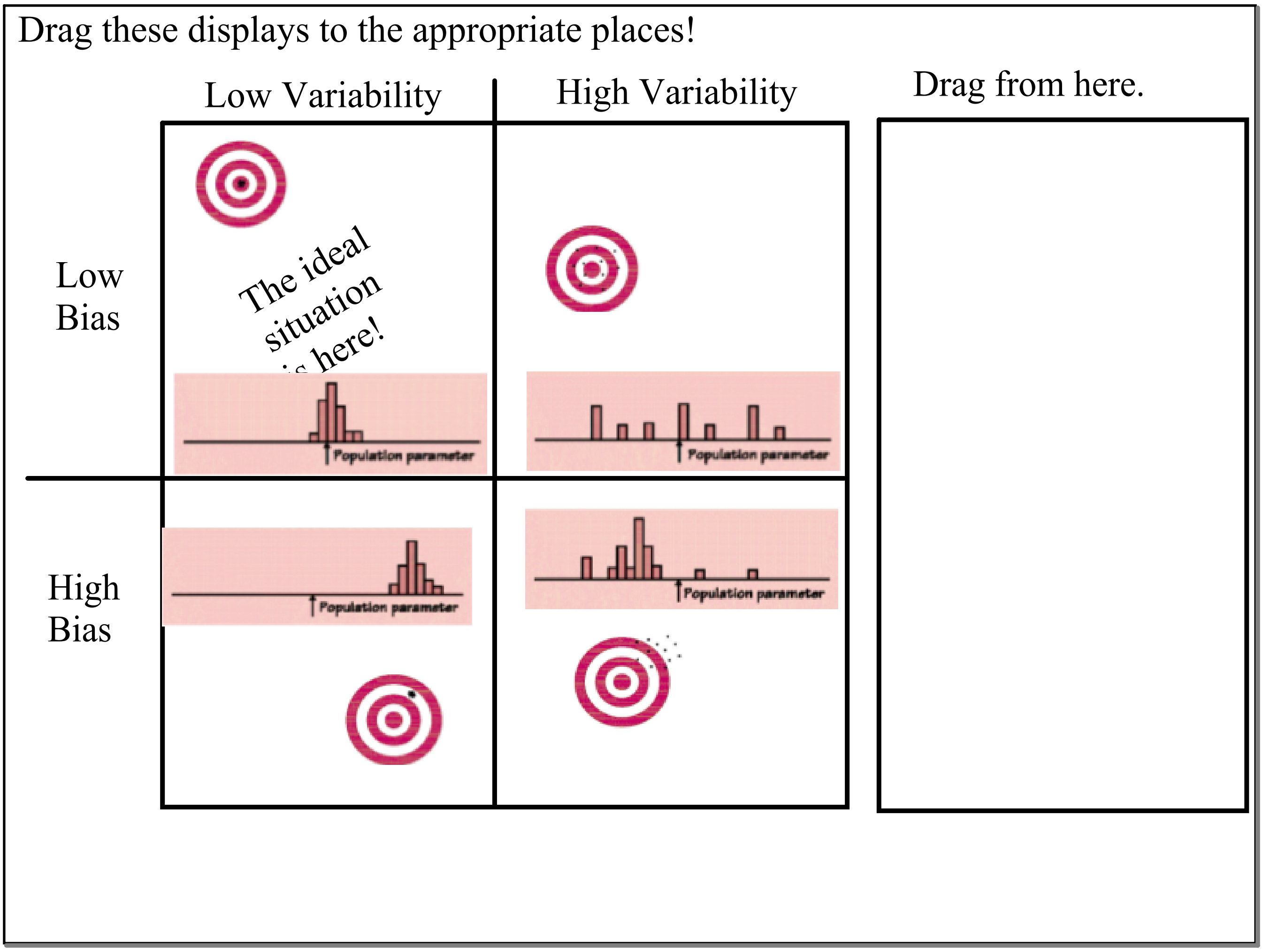
...but that's not a big deal because these statistics will have a distribution with a predictable mean and standard deviation.

The shape of this distribution will be symmetric and approximately normal.
There will be few if any outliers or other deviations from the overall pattern.
The center of this distribution will be close to the parameter.
The spread of this distribution will decrease as n increases.

A statistic (used to estimate a parameter) is **unbiased** if the mean of its sampling distribution is equal to the value of the parameter being estimated.

The **variability of a statistic** is described by the spread of its sampling distribution.

Larger samples give smaller spread. However, as long as the population is much larger than the sample (say, about 10 times larger, $N>10n$), the spread of the sampling distribution is approximately the same for any population size.


Drag these displays to the appropriate places!
Low Variability
High Variability
Drag from here.
Low Bias
High Bias
The ideal
situation
here!
Population parameter
Population parameter
Population parameter
Population parameter

When a SRS is taken, the observed sample statistic is likely to differ from the expected population parameter. This difference is due to:

- Bias from the sampling procedure (we do everything we can to reduce or eliminate this)

  and/or

- Chance error, since every possible subset of the population has the same probability of being chosen

  and/or

- Something significant, leading us to believe that the expected population parameter differs from what we assumed.