Sampling Distribution

Date:

Essential Question: What is a sampling distribution?

Questions:

Notes:

I would like to know the average GPA of all students at Basha High School. Instead of gathering data on every student, I could select a sample and use the sample to estimate the average for all students. Let's say I select a simple random sample of 50 students. I calculate the average GPA for the sample; $\bar{x} = 2.95$.

A **parameter** is a number that describes some characteristic of the population. In statistical practice, the value of a parameter is usually not known because we cannot examine the entire population. Parameters are identified by using Greek letters.
example: $\mu = mean,\ \sigma = standard\ deviation,\ \sigma^2 = variance,\ \rho = proportion$

A **statistic** is a number that describes some characteristic of a sample. The value of a statistic can be computed directly from the sample data. We often use a statistic to estimate an unknown parameter. Statistics are identified by using Roman letters.
example: $\bar{x} = mean,\ s_x = standard\ deviation, {s_x}^2 = variance,\ \hat{p} = proportion$

The value of $\bar{x}$ is considered a statistic because it is describing the mean of a sample. We use μ, a parameter, to describe the mean of the population. Seldom do we actually know the characteristics of a population. Because we are going to use statistics to predict the parameters, we need to know how reliable those values are.

Sampling Variability

How can $\bar{x}$, based on a sample of only a few of the 2600 students, be an accurate estimate of μ? After all, a second random sample taken at the same time would choose different households and likely produce a different value of $\bar{x}$. This basic fact is called sampling variablility; the value of a statistic varies in repeated random sampling.

To make sense of sampling variability, we ask, "What would happen if we took many samples?" Here's how to answer that question.
Take a large number of samples from the same population.
Calculate the statistic (like the sample mean $\bar{x}$ or sample proportion $\hat{p}$) for each sample
Make a graph of the values of the statistics.
Examine the distribution displayed in the graph for shape, center, and spread, as well as outliers or other deviations.

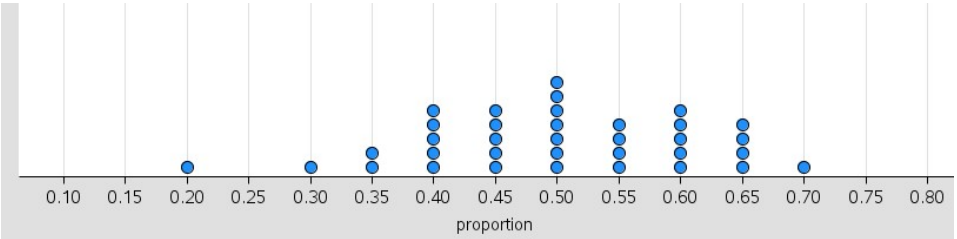
We take a container with red and blue poker chips. We are going to take a sample of 20 chips and record the percentage that are red. We place the chips back into the container, mix them, and then select another sample. Do this 35 times. We will record the percentage of blue chips for each sample and make a dotplot.

Summary:

Date:

Questions:

Notes:



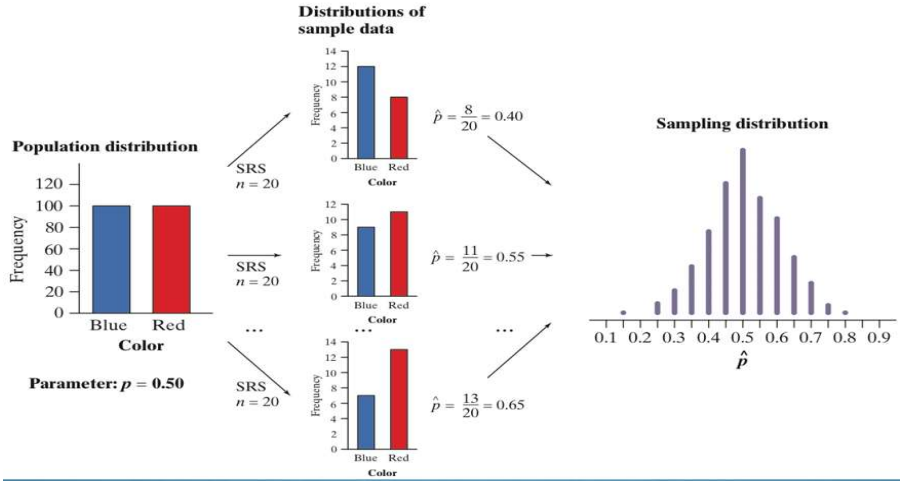
When this was performed in Mr. Hansen's class, this is what we found.
Shape: the graph is roughly symmetric with a single peak at 0.5.
Center: the mean of our sample proportion is 0.499. This is the balance point of the distribution.
Spread: the standard deviation of our sample proportions is 0.112. On average, the values of $\hat{p}$ are about 0.112 away from the mean.
Outliers: There are no obvious outliers or other unusual features.

Of course, the class only took 35 different simple random samples of 20 chips. There are many, many possible SRSs of size 20 from a population of size 200 (about 1.6×10^{27}, actually). If we took every one of those possible samples, calculated the value of $\hat{p}$ for each, and graphed all those $\hat{p}$-values, then we'd have a sampling distribution.

The *sampling distribution* of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population. In practice, it's usually too difficult to take all possible samples of size n to obtain the actual sampling distribution of a statistic. Instead, we can use simulation to imitate the process of taking many, many samples.

example: Container of Chips

We will use a container with 100 red chips and 100 blue chips; this is our population. We will be selecting a sample of 20 chips.



Summary:

Date:

Questions:

Notes:

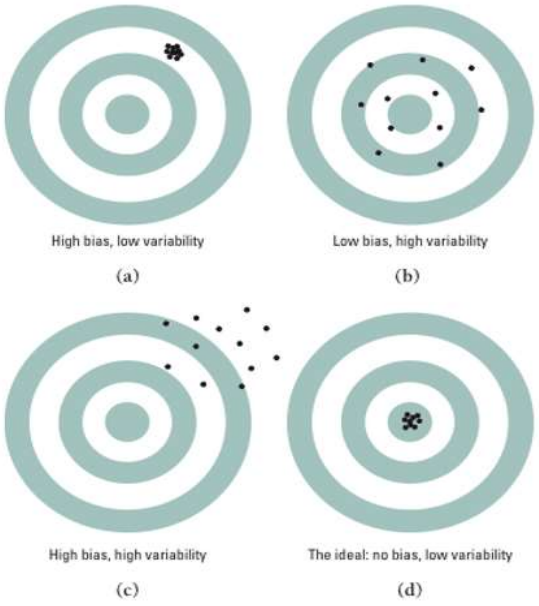
The distribution of sample data shows the values of the variable "color" for the individuals in the sample. For each sample, we record a value for the statistic $\hat{p}$, the sample proportion of red chips. Finally, we collect the values of $\hat{p}$ from all possible samples of the same size and display them in the sampling distribution.

Be careful: The *population distribution* and the *distribution of sample data* describe individuals. A *sampling distribution* describes how a statistic varies in many samples from the population.

****AP EXAM TIP:** Terminology matters. Don't say "sample distribution" when you mean sampling distribution. You will lose credit on free-response questions for misusing statistical terms.

When the mean of the sampling distribution is equal to the parameter we are looking at, then we say that the statistic is unbiased. In the previous example we had a population proportion $\rho = .5$ for blue chips. We found after taking many samples that the sampling distribution had a mean of approximately 0.5; ($\mu_{\hat{p}} = 0.5$). Therefore, we would say that the statistic for proportion $\hat{p}$ is an unbiased estimator for the population proportion (the parameter) ρ.

Keep in mind, that an unbiased estimator does not mean that you will always have the exact value of the parameter, but you will be close. So our goal is to reduce variability and bias. Think of it like shooting at a target. The bull's-eye represents the actual parameter's value.



Bias means that our aim is off and we consistently miss the bull's-eye in the same direction. Our sample values do not center on the population value.

High *variability* means that repeated shots are widely scattered on the target. Repeated samples do not give very similar results.

Our goal is to reduce bias and variability like letter (d).

Summary: