(8-2016)

Section 7.1 - What is a Sampling Distribution? (pp. 420-439)

1. Parameters and Statistics

Hyena Problem - In the hyena problem, we took repeated samples from a population to try to determine the proportion of male hyenas in the population. The proportions you found in the samples are called **statistics** because they describe **samples**. The actual proportion in the population is called a **parameter** because it is from the **population**.

Parameter - 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.

Statistic - 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 statistics to estimate an unknown parameter.

It is essential from this point on that we always distinguish parameters and statistics (p=P, s=S). For example, μ is the population mean so it is a parameter while $\bar{x}$ is a statistic because it is the sample mean. For proportions, p is the population proportion (a parameter) while $\hat{p}$ is the sample proportion (a statistic).

Example - A pediatrician wants to know the 75th percentile for the distribution of heights of 10-year-old boys, so she takes a sample of 50 patients and calculates $Q_3 = 56$ inches.

What is the population? ALL 10 YO BOYS

What is the parameter? 75th %ILE

What is the sample? 50 10 YO BOYS

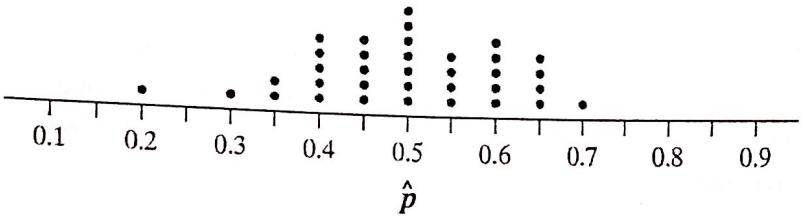
What is the statistic? $Q_3 = 56$ INCHES.

---

2. Sampling Variability - As we saw in the hyena problem, the various samples for any given population produced different sample proportions. This basic fact is called **sampling variability**. In the hyena experiment, we:

- Took repeated samples from the same population,
- Calculated the sample proportion $\hat{p}$ for each sample,
- Made a graph of the values of the statistic,
- Examined the distribution displayed in the graph for shape, center, and spread, as well as outliers and other deviations.

Suppose one group in the hyena experiment took 35 samples of size 20 and their results are shown in the dotplot below.



SOCS:

Shape: ROUGHLY SYMMETRIC; UNIMODAL W/PEAK @ 0.5

Center: MEAN = 0.499 (BALANCE POINT)

Spread: SD = 0.112: ON AVG, VALUES OF $\hat{p}$ ARE .112 FROM MEAN OF 0.499

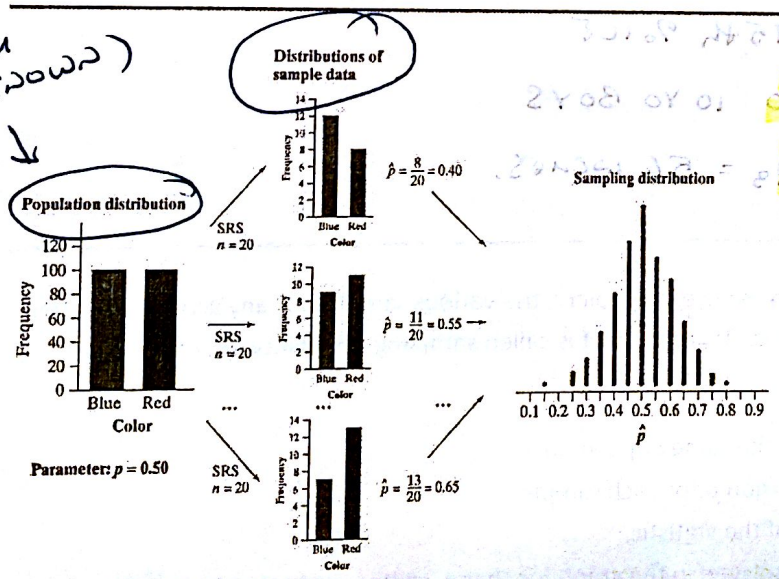
Outliers: NO OUTLIERS OR UNUSUAL FEATURES.

Of course this group only took 35 different simple random samples of 20 hyenas. There are many, many possible SRSs of size 20 from a population of 100. If we took every one of those samples, calculated $\hat{p}$ for each, and graphed all those $\hat{p}$-values, we would have a ***sampling distribution***. $\binom{100}{20} = 5.36 \times 10^{20}$

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.

The sampling distribution is an ideal pattern that would emerge if we looked at *all* possible samples from a given population. (THEORY)

(TRUTH (NOT KNOWN))



There are actually *three* different distributions involved when we sample repeatedly and measure a variable of interest:

(1) The **population distribution** (UNKNOWN)

(2) The **distributions of sample data**

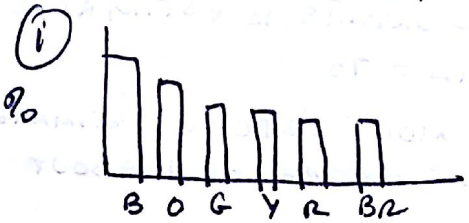
(3) The **sampling distribution**

It is imperative that you can keep these three distributions straight. The population distribution and the distributions of sample data describe *individuals*. The sampling distribution describes how a statistic varies in many samples from the population.

BLUE .24 ORANGE .2 GR .16 YEL .14 R .13 BR .13

Check Your Understanding - Complete CYU on p. 428.

(1)



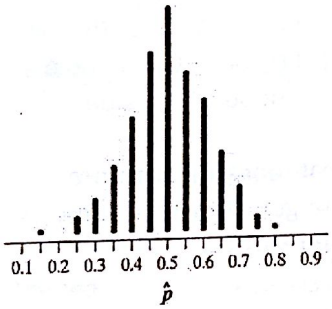
(2) GRAPH OF SRS OF 50 – STUDENTS CHOICE

(3) MIDDLE GRAPH SINCE CENTERED @ $\hat{p} = 0.20$.

IND: M+M'S
VAR: COLORS
PARAMETER: P ORANGE

3. Describing Sampling Distributions

a. Center: Biased and unbiased estimators



How well does the sample proportion of hyenas estimate the population proportion of hyenas? The dotplot shows the approximate sampling distribution of $\hat{p}$. We noted earlier that the center of this distribution is very close to 0.5, the parameter value. In fact, if we took all possible samples of 20 hyenas from the population, calculated $\hat{p}$ for each sample, and then found the mean of those $\hat{p}$-values, we would get exactly 0.5. For this reason, we say $\hat{p}$ is an **unbiased estimator** of p.

Unbiased Estimator - A statistic used to estimate a parameter is an unbiased estimator if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

Note: unbiased does not mean perfect. An unbiased estimator will almost always provide an estimate that is not equal to the population parameter. It is called unbiased because in repeated samples, the estimates will not consistently be too high or too low.

CONSISTENT

Biased Estimator - A statistic used to estimate a parameter is a biased estimator if the mean of its sampling distribution not equal to the true value of the parameter being estimated.

b. Spread: Low variability is better! - To get a trustworthy estimate of an unknown population parameter, start by using a statistic that is an unbiased estimator. This ensures that you do not get an over or underestimate. However, this does not guarantee that the value of the statistic from your sample will be close to the actual parameter value.

The key to success is larger samples. The size of the random sample drives the variability of the sampling distribution. The variability is not a function of the size of the population.

Variability of a Statistic

The variability of a statistic is described by the spread of its sampling distribution. This spread is determined primarily by the *size of the random sample*. Larger samples give smaller spread. The spread of the sampling distribution does not depend on the size of the population, **as long as the population is 10 times larger than the sample** (10% rule).

Check Your Understanding - Complete CYU on p. 434.

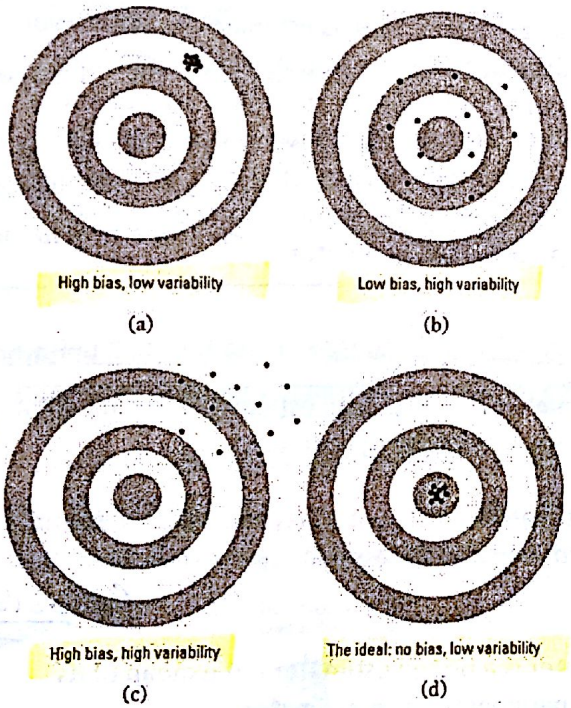
(1) MEDIAN DOES NOT APPEAR TO BE AN UNBIASED ESTIMATOR
MEAN OF MEDIANS = 73.5 POP. MEDIAN = 75

(2) SMALLER. LARGER SAMPLES PROVIDE MORE PRECISE ESTIMATES BECAUSE LARGER SAMPLES INCLUDE MORE INFORMATION ABOUT THE POP. DISTR.

(3) SKEWED LEFT. UNIMODAL.

---

c. Bias, variability and shape - The true value of a population parameter can be thought of as the bull's eye on a target and the sample statistic as the bullet fired at the target. Both bias and variability describe what happens when we take many shots at the target.



Bias means 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.

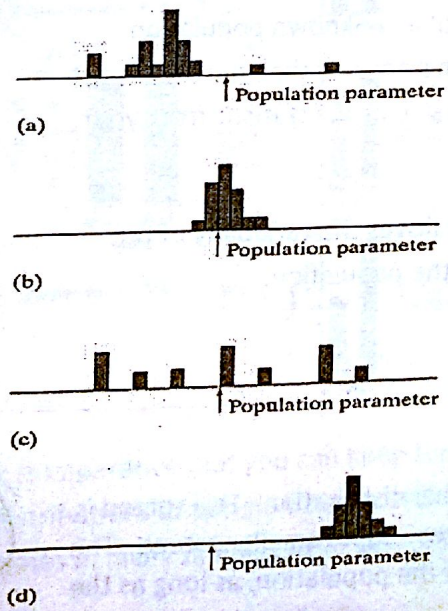
High variability means that repeated shots are widely scattered on the target. Repeated samples are not giving very similar results.

Notice that low variability can accompany high bias and low or no bias can accompany high variability.

Ideally, we would like our estimates to be *accurate* (unbiased) and *precise* (have low variability).

The bottom line is choose a statistic that has low or no bias and minimum variability.

P. 438 #19.



Application - The figure to the left shows histograms of four sampling distributions of different statistics intended to estimate the same parameter.

(a) Which statistics are unbiased estimators? Why?

B AND C BECAUSE MEANS OF SAMPLING DISTR'S = POP. MEANS.

(b) Which statistic does the best job of estimating the parameter? Why?

B. IT IS UNBIASED AND HAS LITTLE VARIABILITY.

HW: Read pp. 420-439; Do problems 1-13 odd, 17, 20, 25*, 26*