To Do

Read Sections 4.1 and 4.2.

Start End-of-Chapter Problems 1-30.

Today's Lecture

(1) Definition of a Point Estimate

(3) Definition of a Point Estimator

(4) Definition of a Sampling Distribution of an Estimator

(5) Review of Results from STAT 230 which can be used to determine Sampling Distributions

(6) The Sampling Distribution of the Sample Mean and Sample Proportion in the Diamond Experiment

Section 4.2: Estimators and Sampling Distributions

Suppose that for a certain study population we are interested in estimating an attribute using observed data $y_1, y_2, \ldots, y_n$.

Suppose also that the attribute of interest can be represented by the parameter θ.

Definition of a Point Estimate

Definition:

A point estimate of θ, is a function

$$\hat{\theta} = g(y_1, y_2, \ldots, y_n)$$

of the observed data $y_1, y_2, \ldots, y_n$ used to estimate the unknown parameter θ.

Example:

For Poisson data with unknown mean θ we use $\hat{\theta} = \bar{y}$ to estimate θ.

Repeated Random Samples

If we take repeated random samples $y_1, y_2, \ldots, y_n$ then the estimates

$$\hat{\theta} = g(y_1, y_2, \ldots, y_n)$$

obtained from the different samples will vary.

For example, the two samples you collected last day did not necessarily have the same sample mean.

Important Idea: Since estimates vary as we take repeated samples, we associate a random variable with these estimates.

Point Estimators

Let $Y_1, Y_2, \ldots, Y_n$ be the potential observations in a random sample.

Associate with the point estimate

$$\hat{\theta} = g(y_1, y_2, \ldots, y_n)$$

the random variable

$$\tilde{\theta} = g(Y_1, Y_2, \ldots, Y_n)$$

Example:

The random variable associated with

$\hat{\theta} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ **is** $\tilde{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Notation

In this course we will use $\hat{\theta}$ to denote an estimate (a numerical value) and $\tilde{\theta}$ to denote the corresponding estimator (a random variable).

This is not notation adopted by all textbooks on statistics but this notation will be useful as you learn to understand the difference between estimates and estimators.

Definition of a Point Estimator

Definition:

A point estimator is a random variable which is a function $\tilde{\theta} = g(Y_1, Y_2, \ldots, Y_n)$ **of the random variables** $Y_1, Y_2, \ldots, Y_n$ **.**

An estimator is simply a rule that tells us how to process the data to obtain an estimate of an unknown parameter θ**.**

The numerical value $\hat{\theta} = g(y_1, y_2, \ldots, y_n)$ **is the value obtained using this rule for a particular observed data set** $y_1, y_2, \ldots, y_n$**.**

Sampling Distribution of an Estimator

Since $\tilde{\theta}$ is a random variable, it has a distribution.

In other words, if $\tilde{\theta}$ is a discrete random variable then it has a probability function and if $\tilde{\theta}$ is a continuous random variable then it has a probability density function.

Sampling Distribution of an Estimator

Definition:

The distribution of an estimator $\tilde{\theta}$ is called the sampling distribution of the estimator.

Sampling Distribution of the Sample Mean in the Diamond Expt.

How would you determine the sampling distribution of the sample mean $\overline{Y}$ in the Diamond Experiment?

Sampling Distribution of the Sample Mean in the Diamond Expt.

Brute Force Method:

List all the possible sets of size 10.

Calculate the sample mean for each set.

Create a table of possible values of the sample mean and the probability of each possible value.

This is the probability function for $\overline{Y}$.

Sampling Distribution of an Estimator

The sampling distribution can sometimes be determined exactly using results you learned in STAT 230.

Sum of Independent Poisson Random Variables

If $X \sim \text{Poisson}(\mu)$ and $Y \sim \text{Poisson}(\lambda)$ independently then
$X + Y \sim \text{Poisson}(\mu + \lambda)$.

Sum of Independent Binomial Random Variables

If $X \sim \text{Binomial}(n,\theta)$ and $Y \sim \text{Binomial}(m,\theta)$ independently then $X + Y \sim \text{Binomial}(n+m,\theta)$.

Linear Combination of Independent Normal Random Variables

Suppose $Y_i \sim N(\mu_i, \sigma_i^2), i = 1,2,\ldots,n$ independently.

Then

$$\sum_{i=1}^{n} a_i Y_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right)$$

Corollary

Suppose $Y_i \sim N(\mu, \sigma^2),\ i = 1,2,\ldots,n$ independently.

Let $S_n = \sum_{i=1}^{n} Y_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Then

$$S_n \sim N(n\mu, n\sigma^2) \quad \text{and} \quad \bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Sampling Distribution of an Estimator

In some situations the sampling distribution must be determined approximately using results like the Central Limit Theorem.

Central Limit Theorem

Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with

$$E(Y_i) = \mu \quad \text{and} \quad Var(Y_i) = \sigma^2, \quad i = 1,2,\ldots,n.$$

Let $$Z_n = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}.$$

For sufficiently large n,

Z_n has approximately a $N(0,1)$ distribution .

Note : Although the limiting Normal distribution does not depend on the distributions of $Y_1, Y_2, \ldots, Y_n$, the rapidity of the approach to normality depends very much on the shapes of these distributions.

Normal Approximation to the Binomial

Suppose Y ~ Binomial(n,θ).

If n is large then by the Central Limit Theorem

$$\frac{\frac{Y}{n}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim G(0,1) \text{ approximately.}$$

Normal Approximation to the Poisson

Suppose $Y \sim \text{Poisson}(\theta)$.

If $\theta \geq 5$ then by the Central Limit Theorem, Y has approximately a $N(\theta, \theta)$ distribution.

Normal Approximation to the Poisson

More generally if Y_i ~ Poisson(θ), i=1,2,…,n independently and if n is large then by the Central Limit Theorem

$$\frac{\bar{Y}-\theta}{\sqrt{\frac{\theta}{n}}} \sim G(0,1) \text{ approximately.}$$

Normal Approximation to the Exponential

If Y_i ~ Exponential(θ), i=1,2,…,n independently then if n is large then by the Central Limit Theorem

$$\frac{\bar{Y}-\theta}{\frac{\theta}{\sqrt{n}}} \sim G(0,1) \text{ approximately.}$$

Sampling Distribution of an Estimator

In more complicated situations the sampling distribution must be determined by computer simulation.

See the video: What is a sampling distribution?

available at www.watstat.ca

Sampling Distribution of an Estimator

The experiment you were doing in Friday's class is an example of a way to determine the approximate sampling distribution.

Problem 3 on Assignment 2 is another example of determining the approximate sampling distribution.

Diamond Experiment

In the Diamond experiment you sampled 10 diamonds (twice).

For each sample of 10 you determined the average of the numbers on the diamonds.

For each sample of 10 you also recorded the proportion of diamonds which were red (Populations 1 and 2) or blue (Populations 3 and 4).

Diamond Experiment

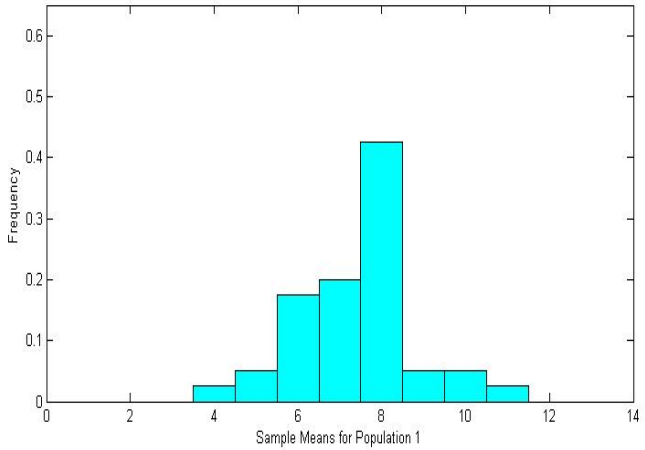
In most cases you obtained different sample means and different sample proportions for your two samples.

Also different students got different sample means and sample proportions.

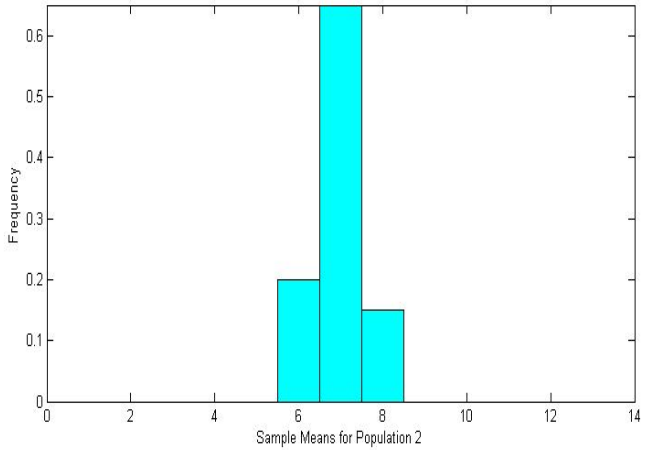
I asked you to draw two samples to illustrate that when you take repeated samples you don’t always get the same sample mean or sample proportion.

Histograms of Observed Means

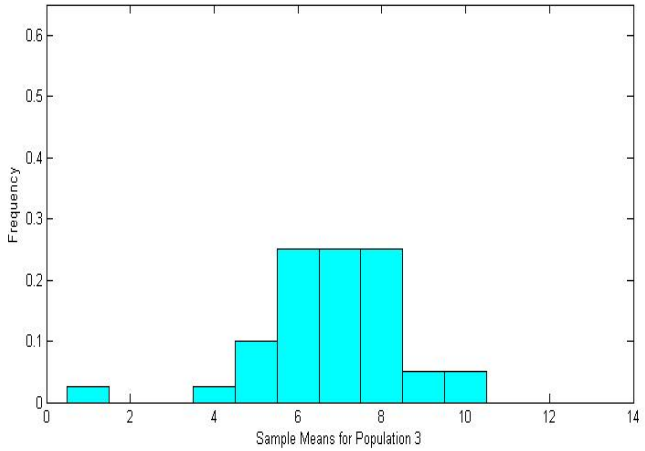
Population 1



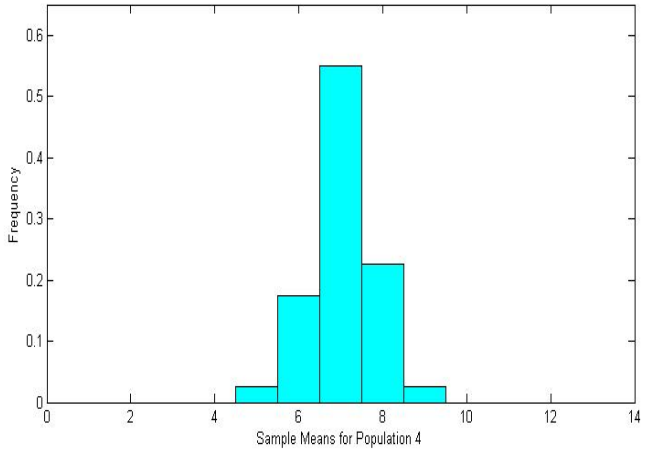
Population 2



Population 3



Population 4



Diamond Experiment

The histograms tell us about the sampling distribution of the random variable

$$\tilde{\mu} = \overline{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i$$

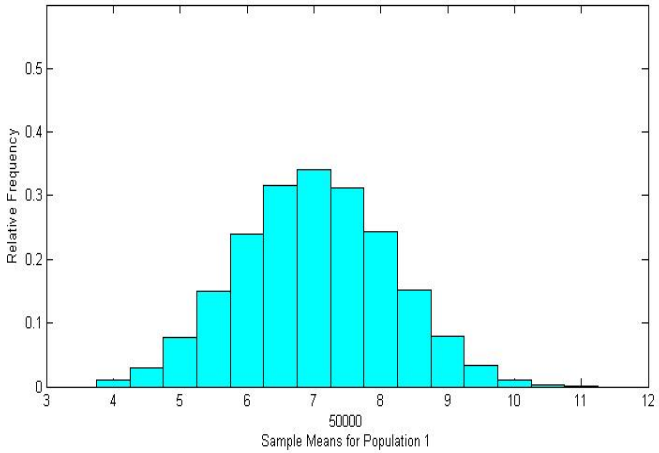
when we draw samples of size 10 repeatedly from a population.

For the data collected in class there were only 40 sample means observed for each population.

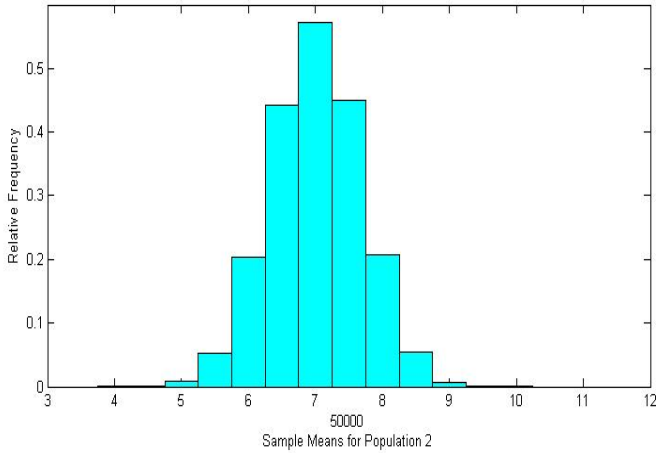
What if we drew 50,000 samples of size 10?

Histograms of 50000 Sample Means

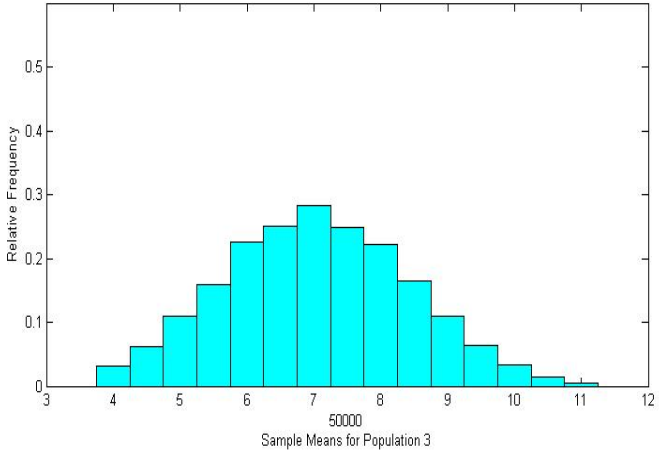
Population 1



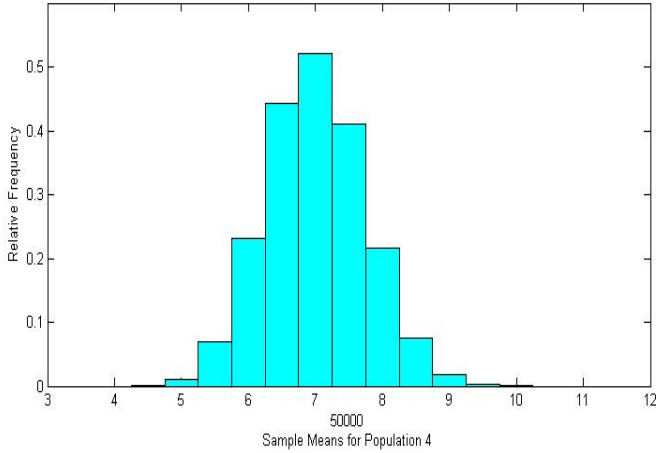
Population 2



Population 3



Population 4



Sampling Distribution of $\bar{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i$

Since 50000 samples of size 10 were drawn from each of the 4 populations, these histograms are a good approximation to the actual sampling distribution (probability function) for $\bar{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i$

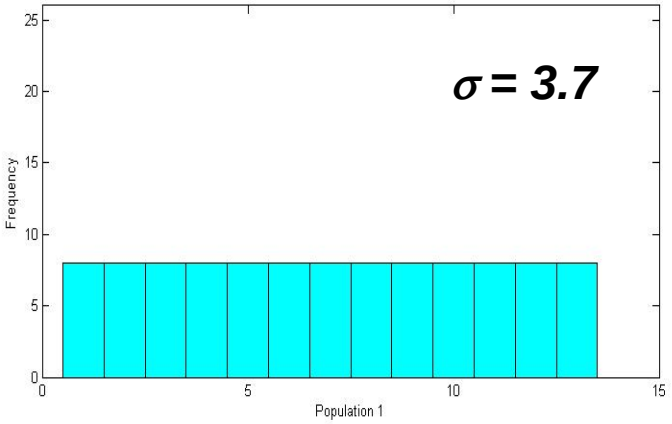
What similarities and differences do you notice?

Does the distribution of $\bar{Y}$ **depend on the population you are sampling from?**

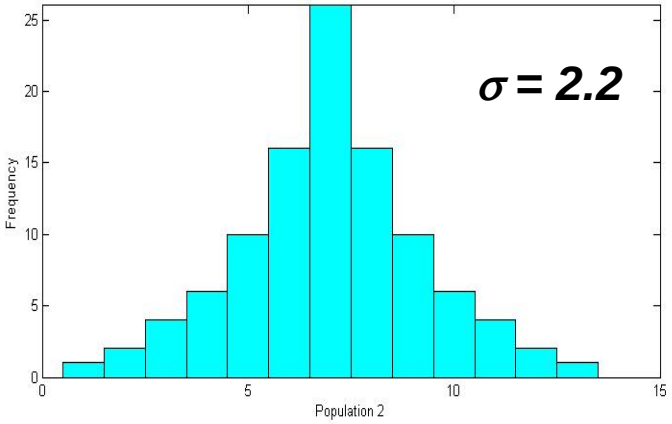
If so how?

Diamond Populations, $\mu = 7$

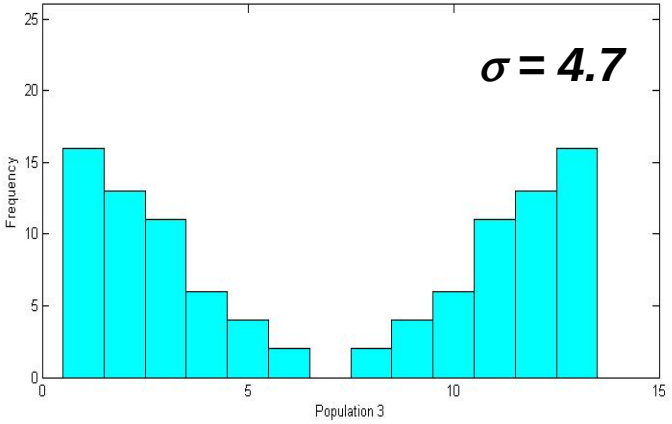
Population 1



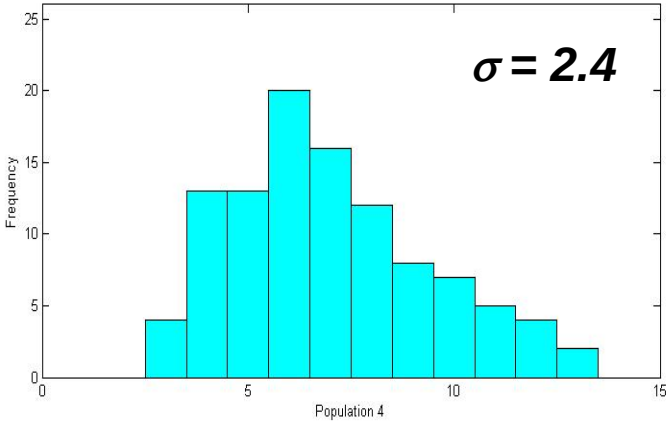
Population 2



Population 3



Population 4



Is the Sample Mean a Good Estimate of the Unknown Population Mean?

Of course we usually only have one sample (not 50000 samples) of observations to estimate the unknown population mean *μ*.

Is $\hat{\mu} = \bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i$ **a good estimate of *μ*? Why?**

Does how good the estimate is depend on the population you are sampling from? Is so in what way?

How can we improve the estimate of *μ*?

Sampling Distribution of $\bar{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i$

$$E(\bar{Y}) = \mu \quad \text{and} \quad sd(\bar{Y}) \approx \frac{\sigma}{\sqrt{n}}$$

where *n* = number of observations, μ is the mean of the population and σ is the standard deviation of the population.

($sd(\bar{Y}) \approx \sigma / \sqrt{n}$ **since sampling was done without replacement.)**

$E(\bar{Y}) = \mu$ **for all values of n.**

As n increases $sd(\bar{Y})$ **decreases.**

How good an estimate is the sample mean of the population mean?

Based on the histograms of 50000 means, how often is the sample mean within 1 unit of the population mean for Population 1?

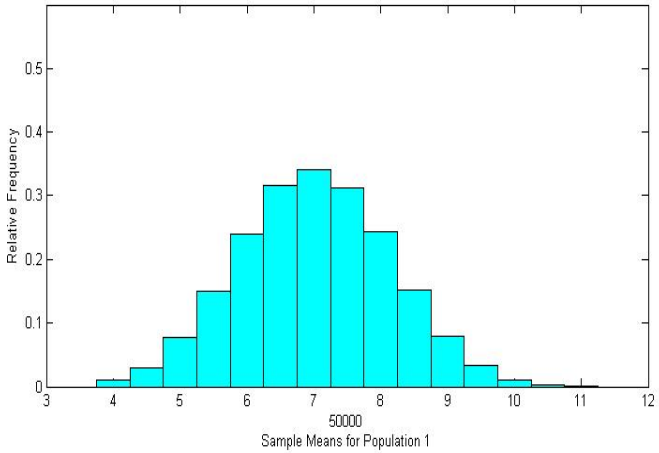
For Population 2?

For Population 3?

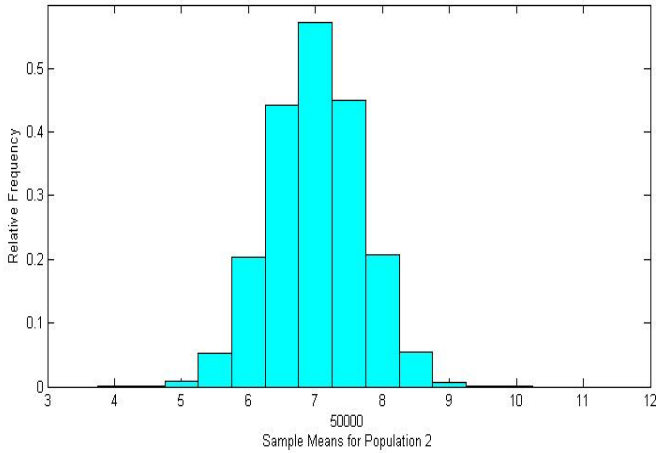
For Population 4?

Histograms of 50000 Sample Means

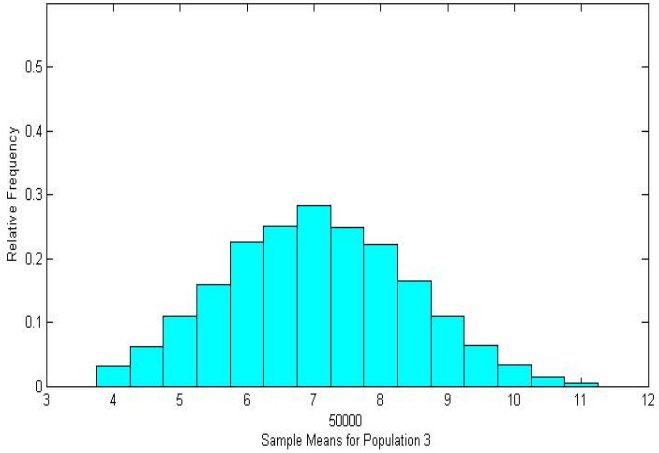
Population 1



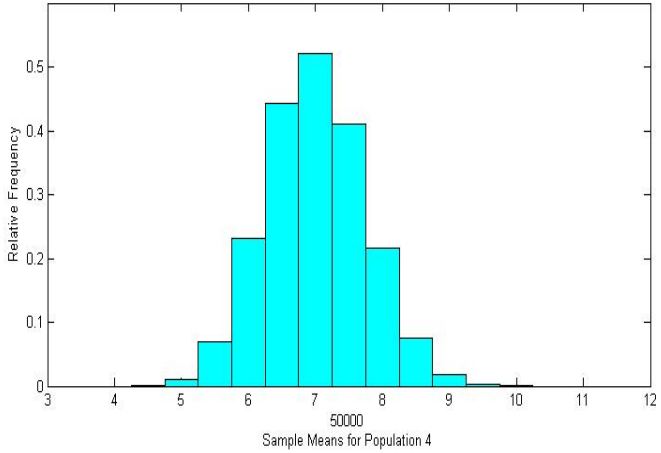
Population 2



Population 3



Population 4



How good an estimate is the sample mean of the population mean?

Based on the histograms of 50000 means, how often is the sample mean within 1 unit of the population mean for Population 1?

73% of the time

For Population 2? **94% of the time**

For Population 3? **61% of the time**

For Population 4? **91% of the time**

What factors affect these proportions?

How good an estimate is the sample mean of the population mean?

Based on the histograms of 50000 means, how often is the sample mean within 1 unit of the population mean for Population 1?

73% of the time

For Population 2? **94% of the time**

For Population 3? **61% of the time**

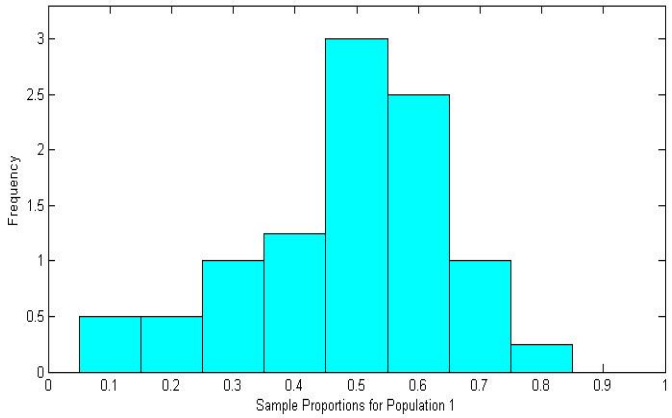
For Population 4? **91% of the time**

What factors affect these values?

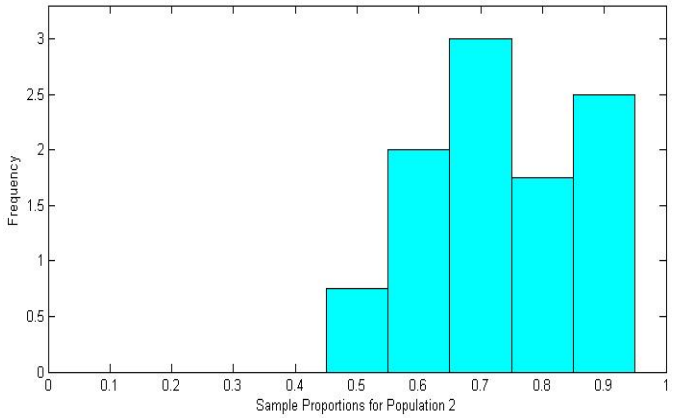
***n* = sample size, σ = the standard deviation of the population, and shape of the population**

Histograms of Observed Proportions

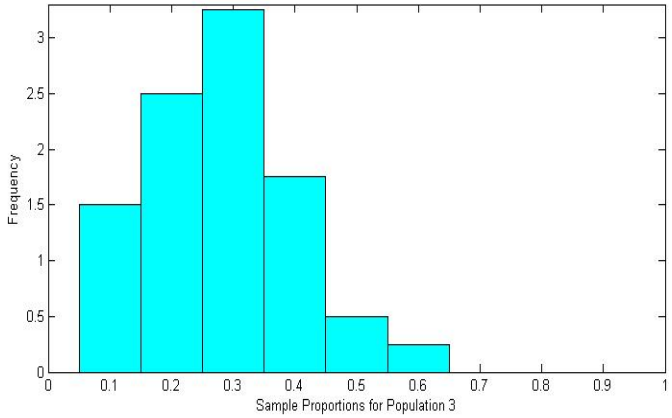
Population 1



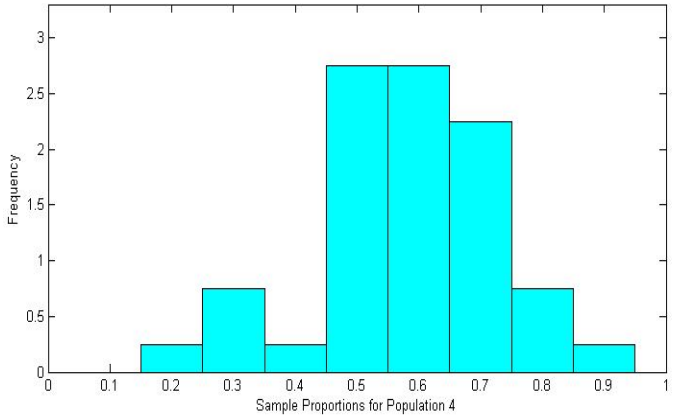
Population 2



Population 3

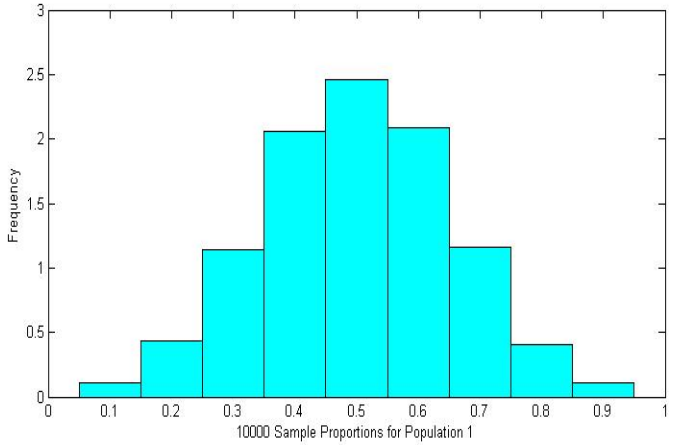


Population 4

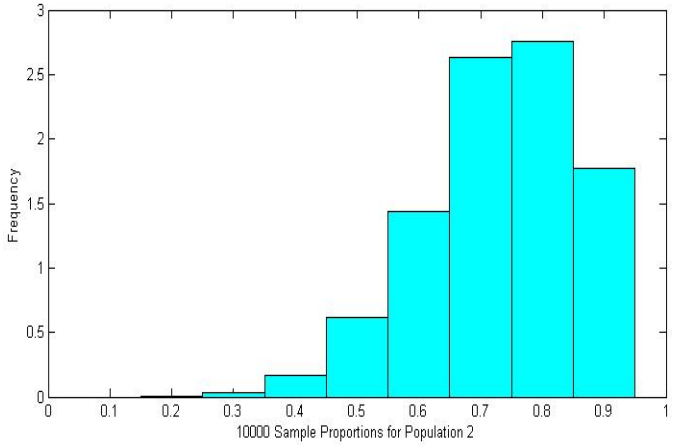


Histograms of 10000 Sample Proportions

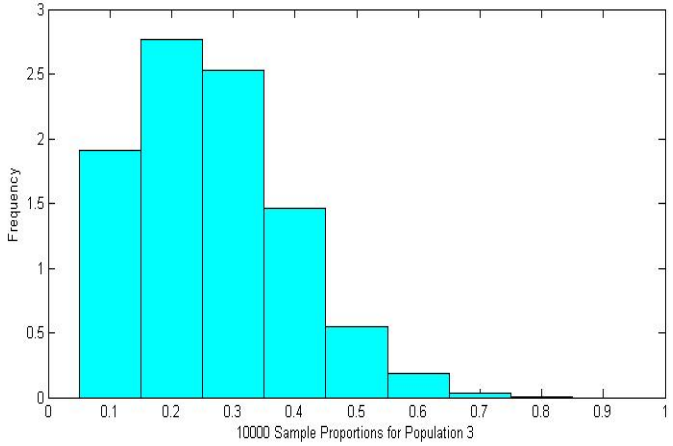
Population 1



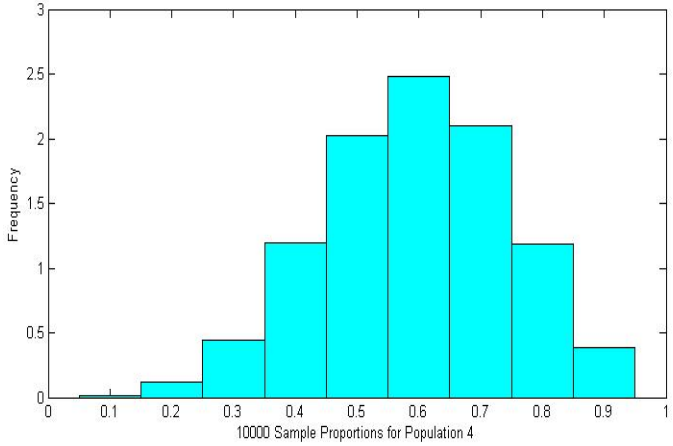
Population 2



Population 3



Population 4



True Population Proportions

Population 1: Proportion Red = 0.5

Population 2: Proportion Red = 0.75

Population 3: Proportion Blue = 0.25

Population 4: Proportion Blue = 62/104
= 0.6

How good an estimate is the sample proportion of the population proportion?

For Population 1 how often is the sample proportion within 0.15 of the population proportion?

For Population 2?

For Population 3?

For Population 4?

How good an estimate is the sample proportion of the population proportion?

For Population 1 how often is the sample proportion within 0.16 of the population proportion? **66% of the time**

For Population 2? **86% of the time**

For Population 3? **86% of the time**

For Population 4? **78% of the time**

What factors affect these values?

How good an estimate is the sample proportion of the population proportion?

For Population 1 how often is the sample proportion within 0.16 of the population proportion? **66% of the time**

For Population 2? **86% of the time**

For Population 3? **86% of the time**

For Population 4? **78% of the time**

What factors affect these values?

***n* = sample size, θ = the population proportion**

Why?

$$E\left(\frac{Y}{n}\right) = \theta$$

and

$$Var\left(\frac{Y}{n}\right) \approx \frac{\theta(1-\theta)}{n}$$

(since sampling without replacement)