



Statistics

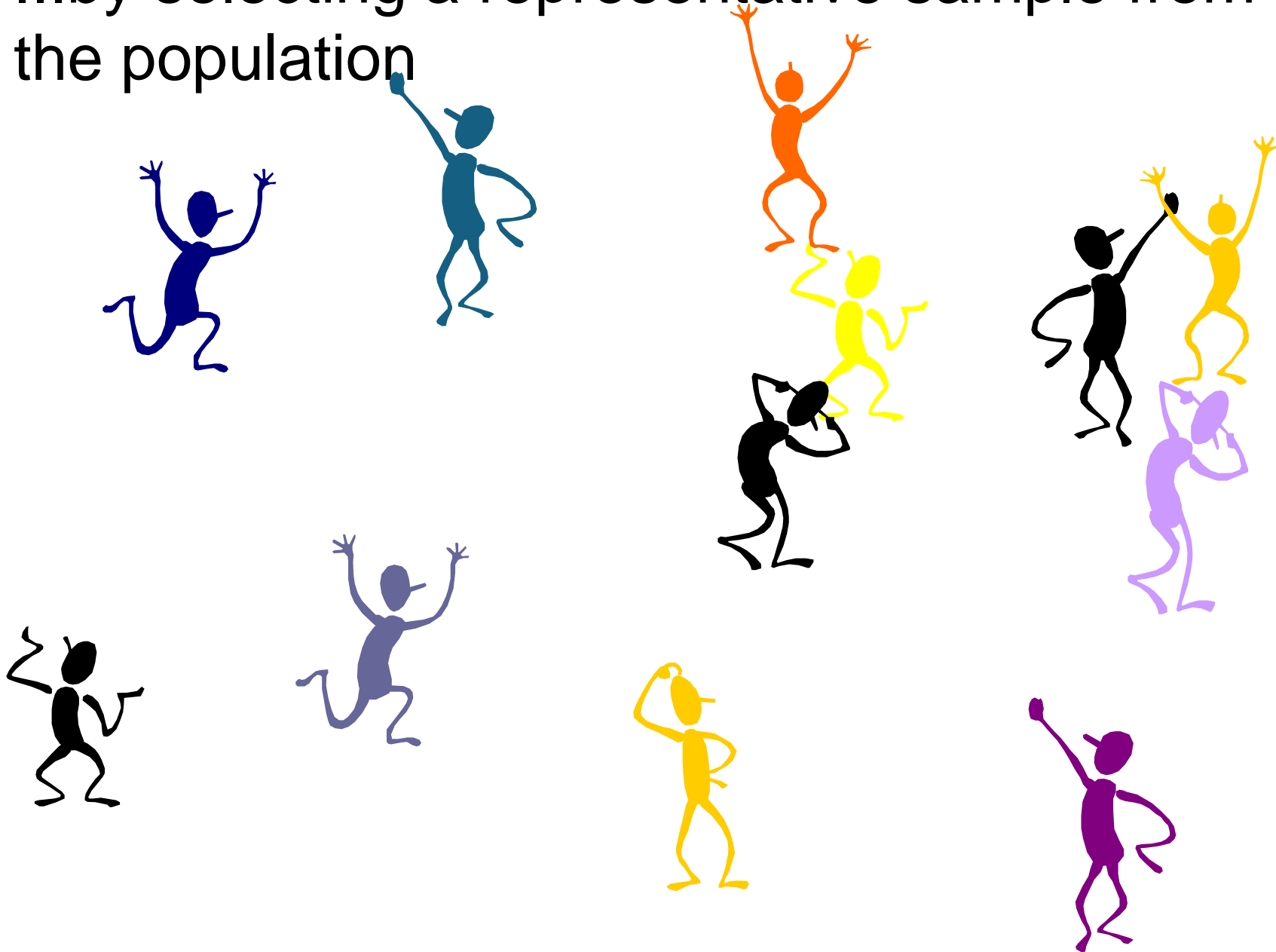
Sampling

- **Population** – A group that includes all the cases (individuals, objects, or groups) in which the researcher is interested.
- **Sample** – A relatively small subset from a population.
- **Simple Random Sample** – A sample designed in such a way as to ensure that
 - (1) every member of the population has an equal chance of being chosen and
 - (2) every combination of N members has an equal chance of being chosen.
- This can be done using a computer, calculator, or a table of random numbers

Population inferences can be made...

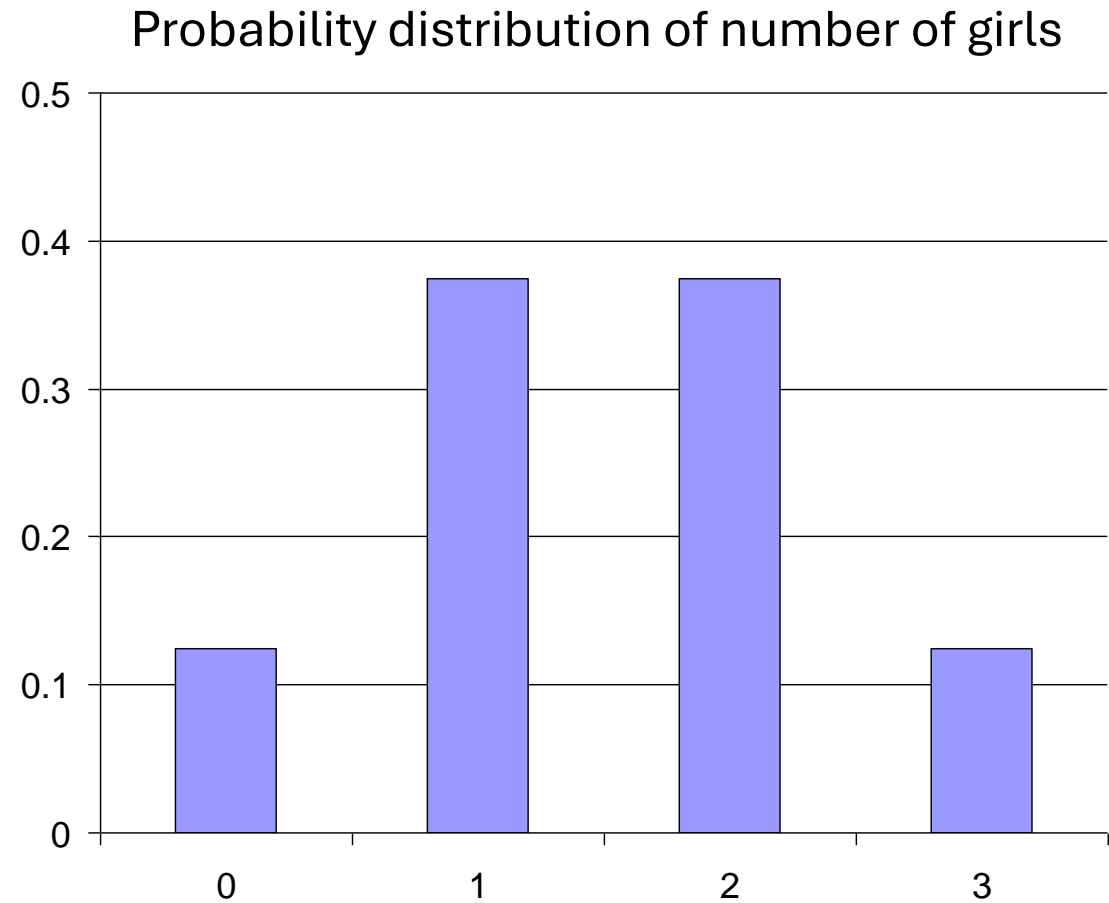


...by selecting a representative sample from the population



How about family of three?

Num Girls	child #1	child #2	child #3
0	B	B	B
1	B	B	G
1	B	G	B
1	G	B	B
2	B	G	G
2	G	B	G
2	G	G	B
3	G	G	G



Probability distributions: Permutations

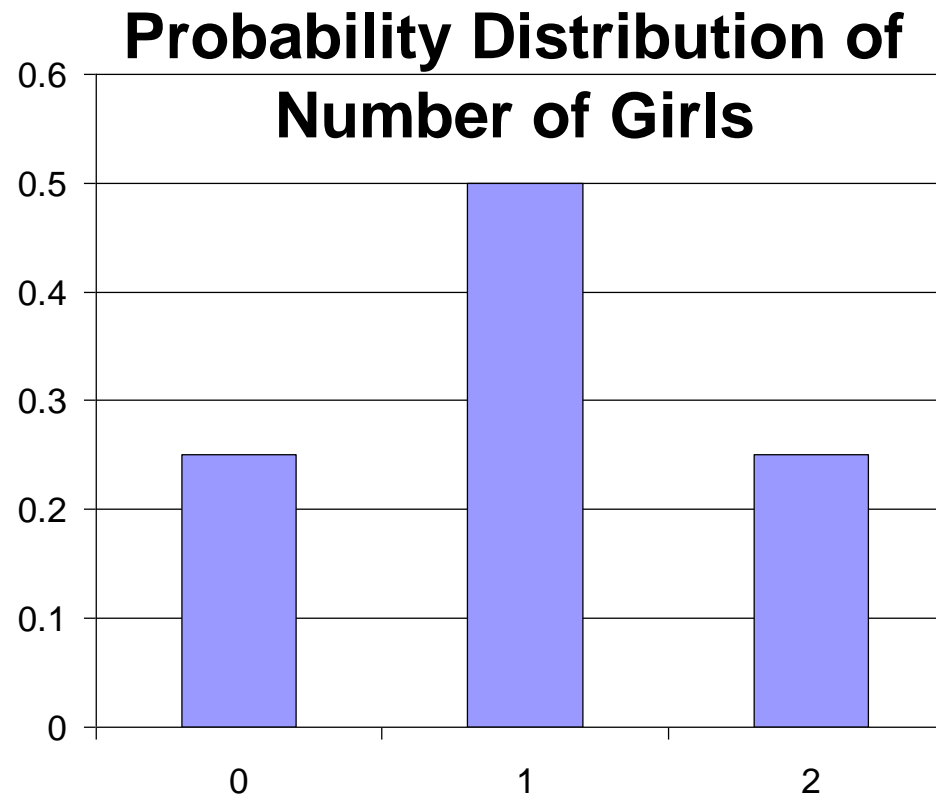
What is the probability distribution of number of girls in families with two children?

2 GG

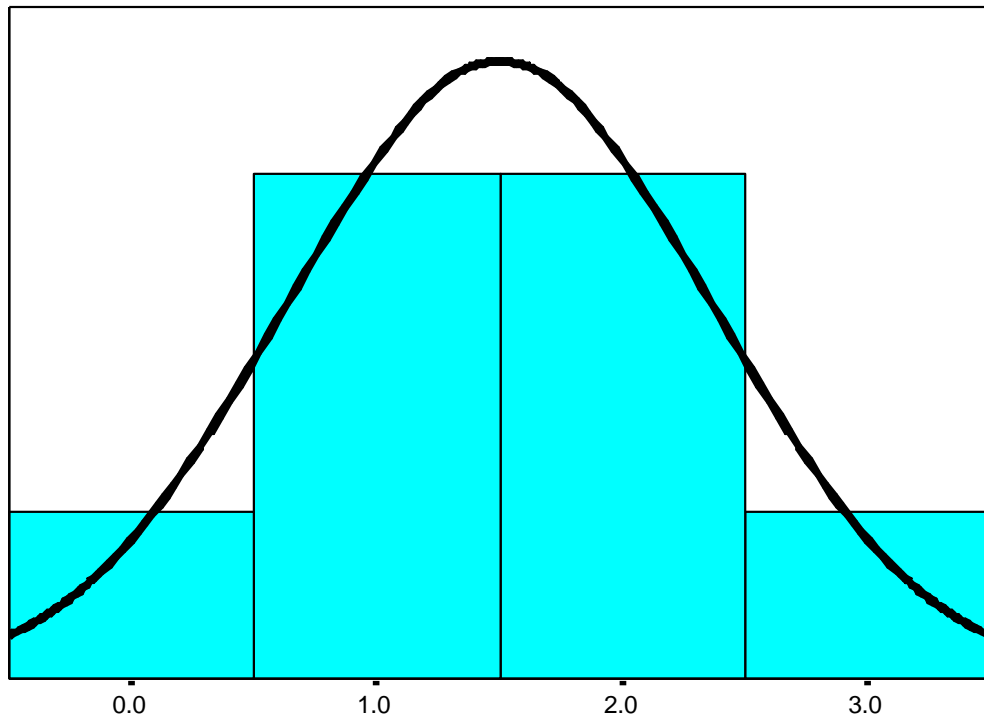
1 BG

1 GB

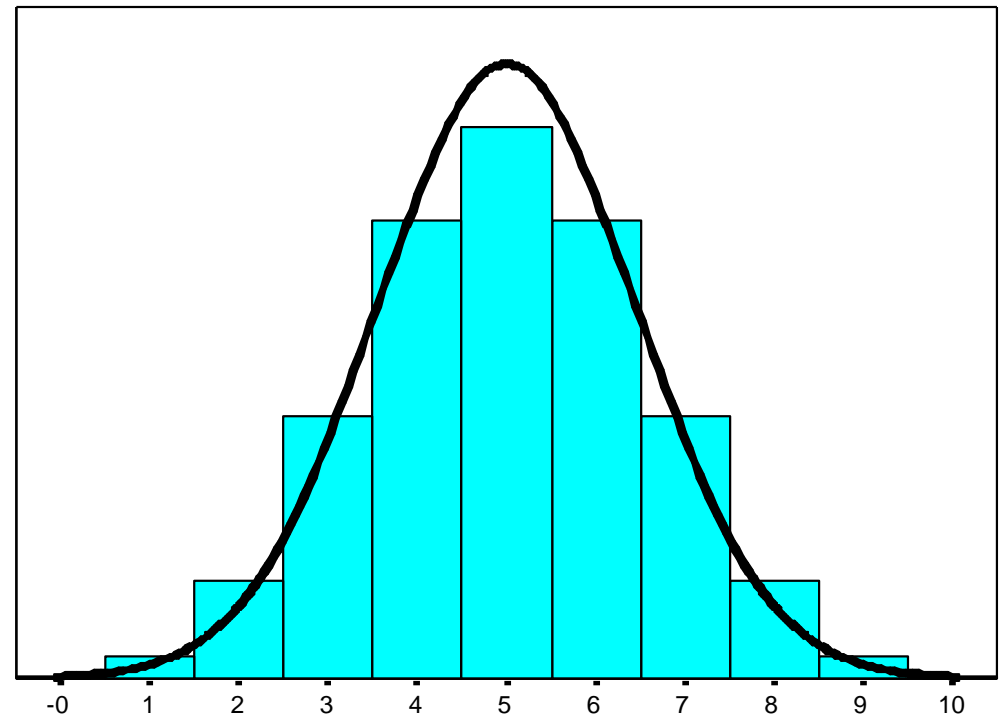
0 BB



As family size increases, the distribution looks more and more normal.



Number of Successes



Number of Successes

Coin toss

- Toss a coin 30 times
- Tabulate results
- Think of the coin tosses as samples of all possible coin tosses

Sampling Distribution

- Imagine repeatedly taking samples of the same size from the large population and calculating a statistic (like the mean or variance) for each sample.
- The probability distribution of these calculated statistics is called the sampling distribution.
- Aim of sampling
 - Reduces cost of research (e.g. political polls)
 - Generalize about a larger population (e.g., benefits of sampling city r/t neighborhood)
 - In some cases (e.g. industrial production) analysis may be destructive, so sampling is needed

Central Limit Theorem

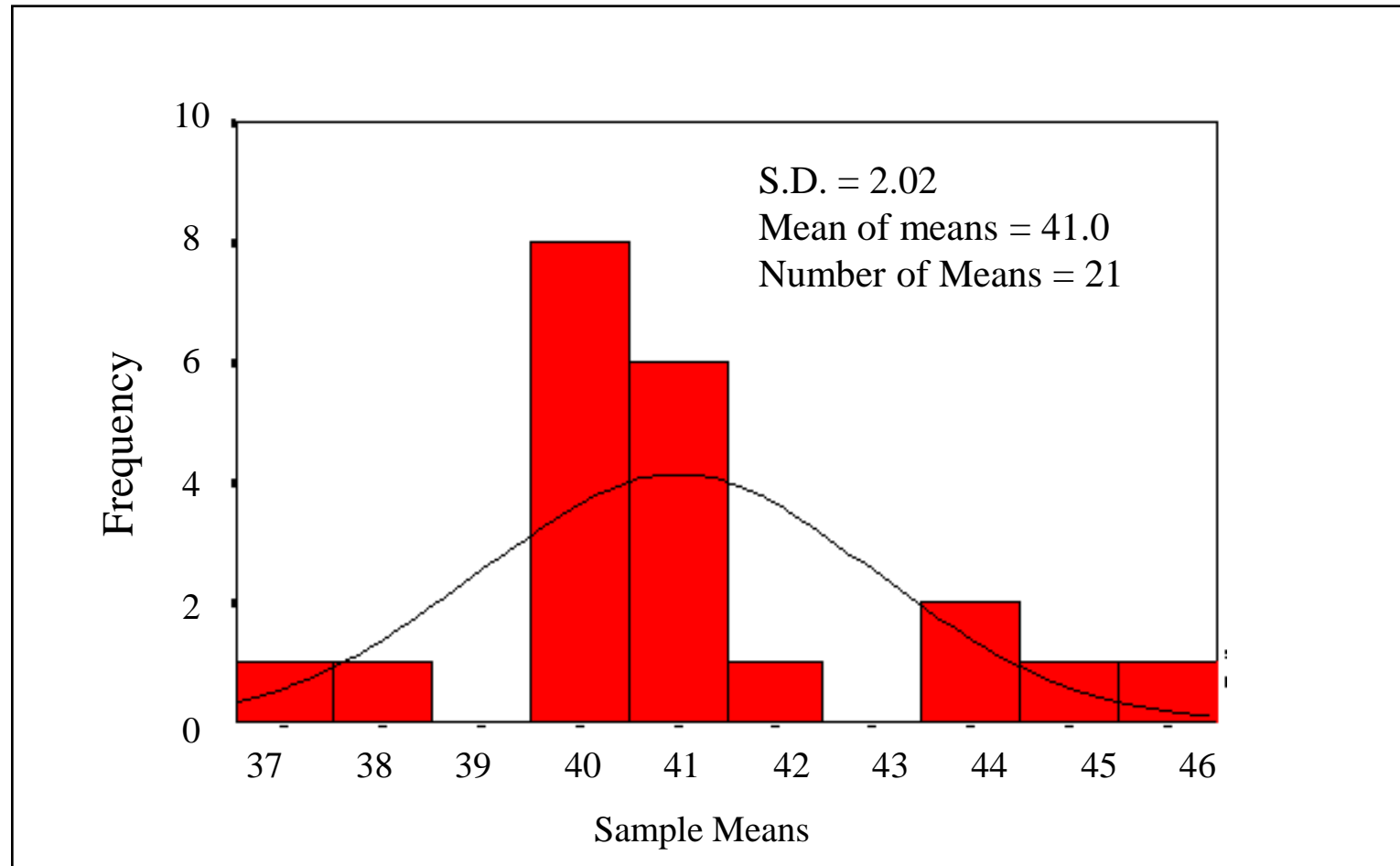
- No matter what we are measuring, the distribution of any measure across all possible samples we could take, approximates a normal distribution, as long as the number of cases in each sample is about 30 or larger.

If we repeatedly drew samples from a population and calculated the mean of a variable or a percentage, those sample means or percentages would be normally distributed.

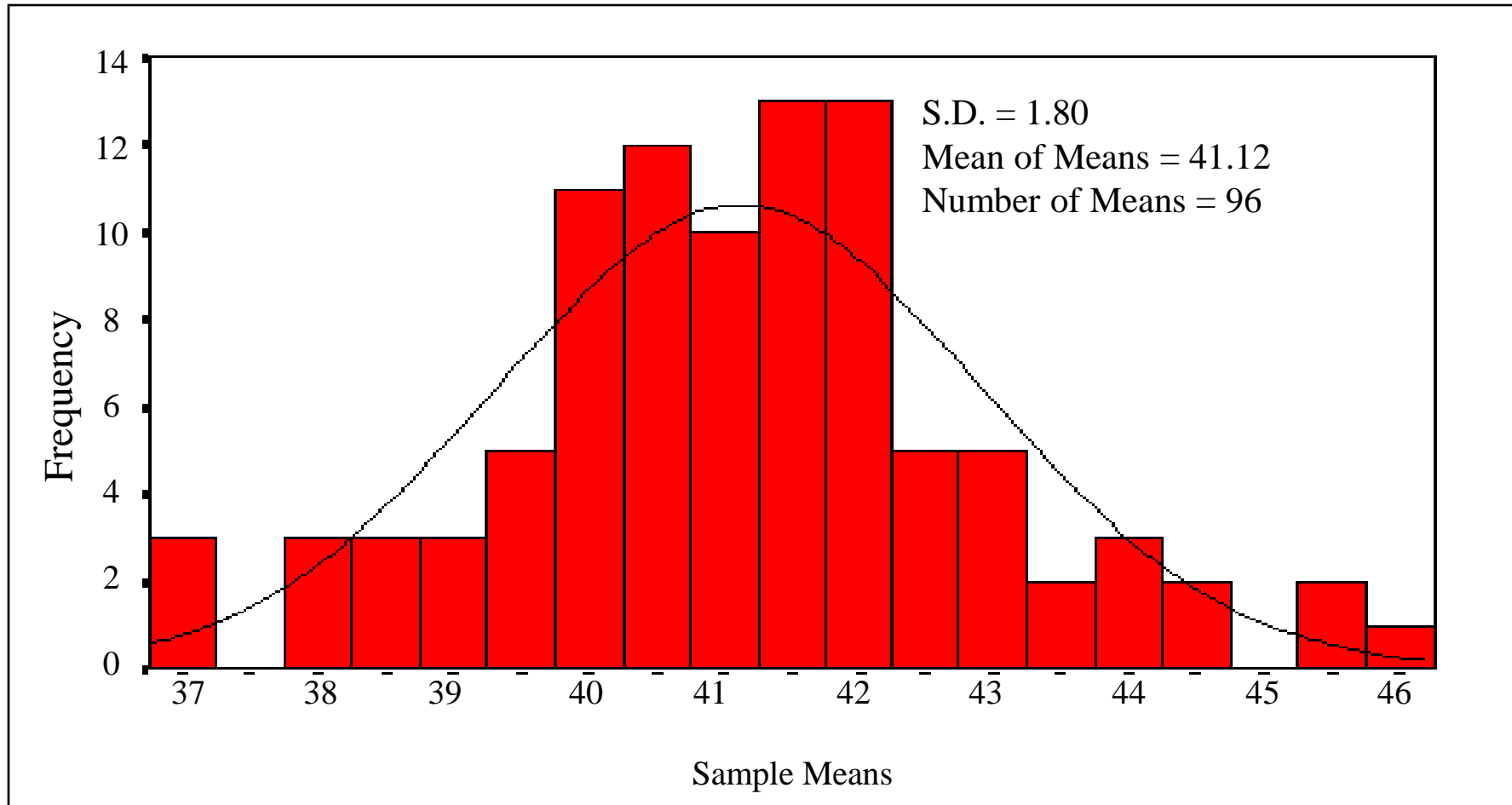
The Mean and Standard Deviation of the Sample Mean

- Suppose we wish to estimate the mean μ of a population. In actual practice we would typically take just one sample.
- Imagine however that we take sample after sample, all with same sample size n , and compute the sample mean \bar{x} each time.
- The sample mean \bar{x} is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty.
- Consider \bar{X} , as a random variable of the sample mean, and write x for the values that it takes.
- The random variable \bar{X} has a mean, denoted $\mu_{\bar{x}}$, and a standard deviation, denoted $\sigma_{\bar{x}}$.

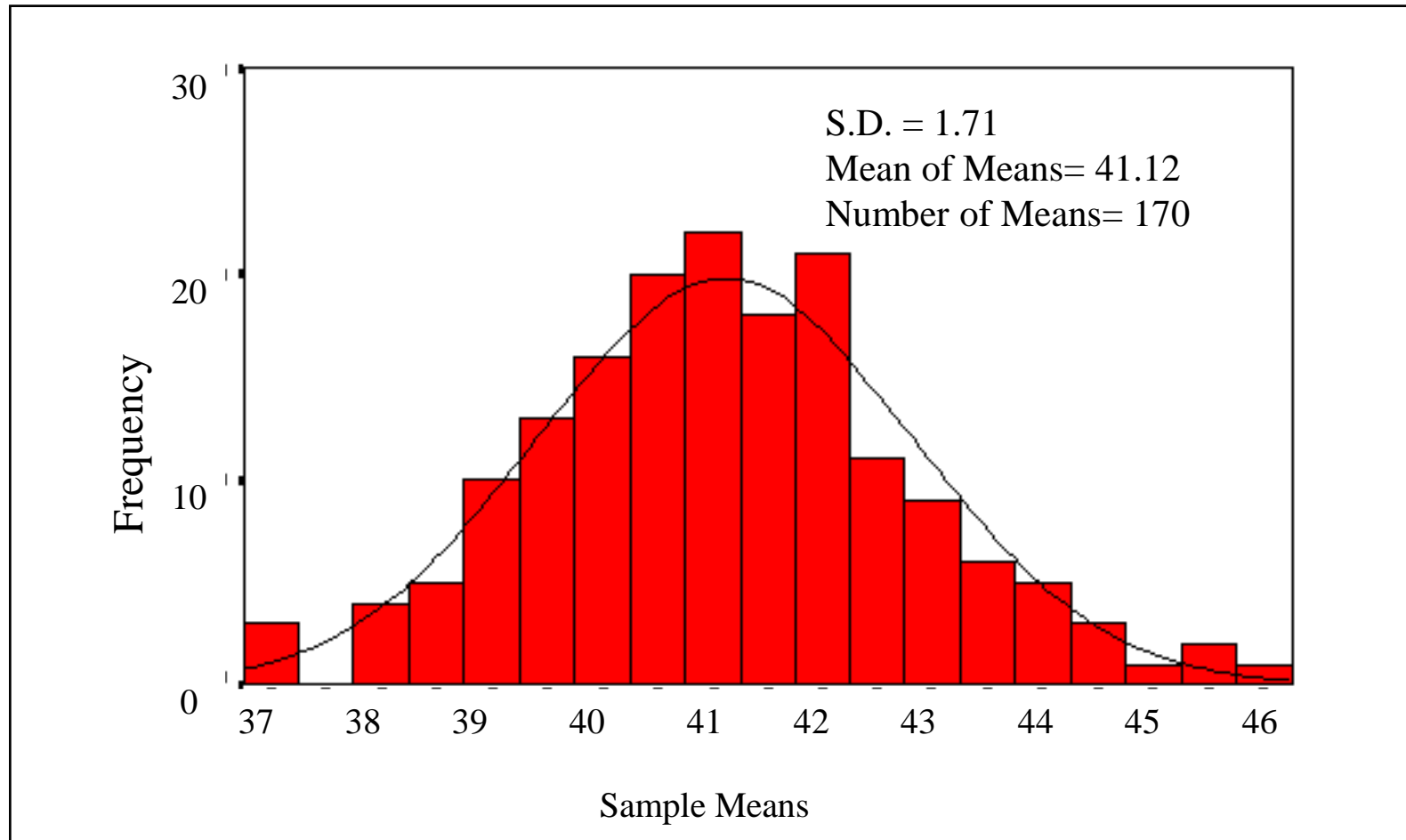
Distribution of Sample Means with 21 Samples



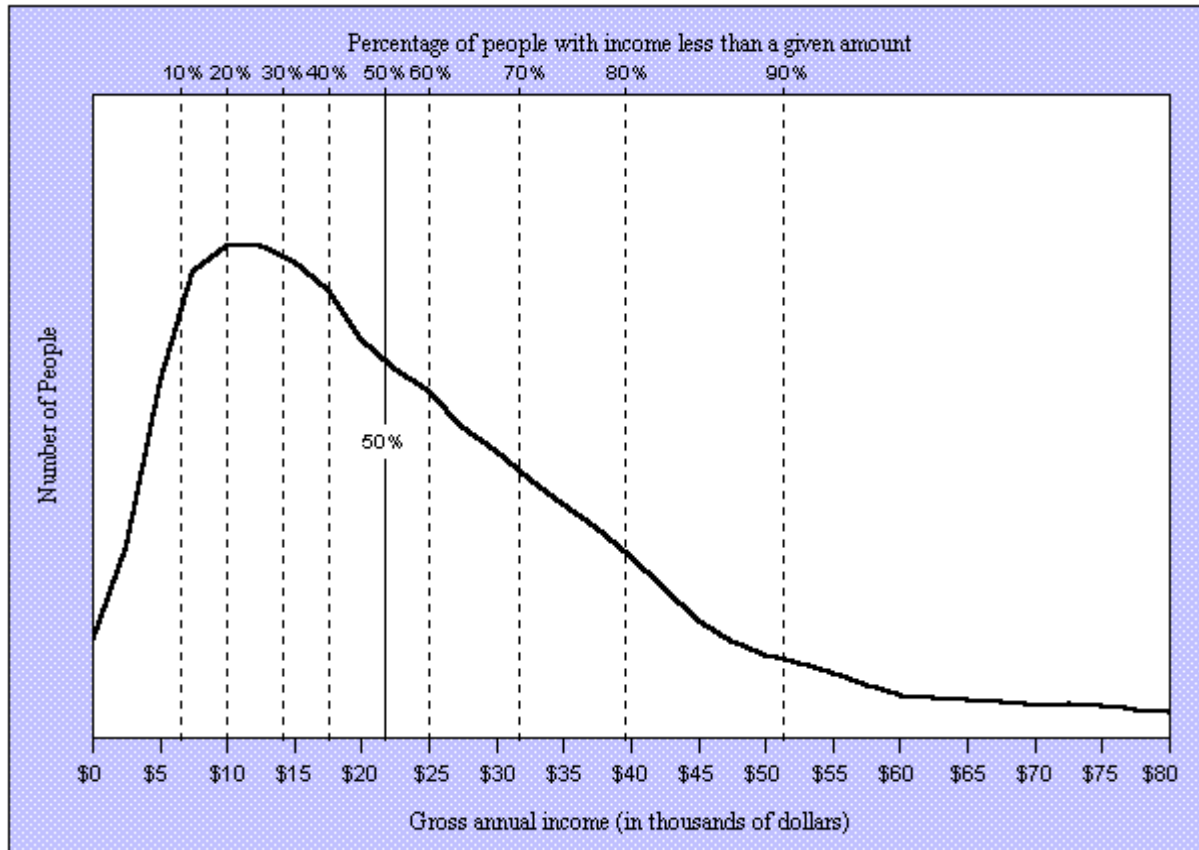
Distribution of Sample Means with 96 Samples



Distribution of Sample Means with 170 Samples

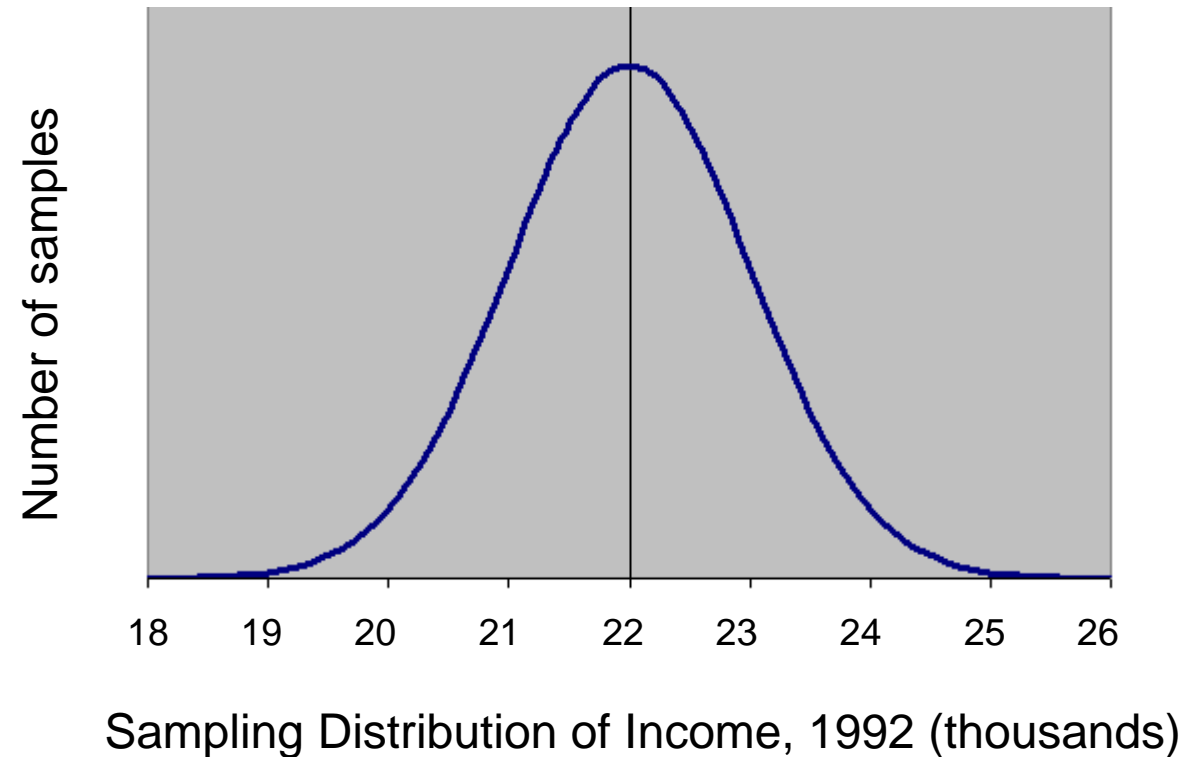


Most empirical distributions are not normal:



U.S. Income distribution 1992

But the sampling distribution of mean income over many samples *is* normal



Central Limit Theorem

When large samples usually greater than thirty are taken into consideration then the distribution of sample arithmetic mean approaches the normal distribution irrespective of the fact that random variables were originally distributed normally or not.

Let us assume we have a random variable X .

Let σ be its standard deviation and μ is the mean of the random variable.

Now as per the Central Limit Theorem, the sample mean \bar{X} will approximate to the normal distribution which is given as $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.

Central Limit Theorem Formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Sample Mean = Population Mean = μ

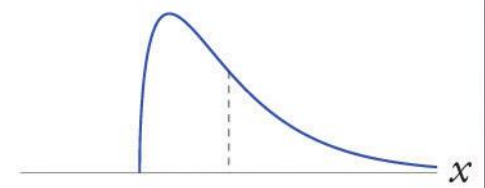
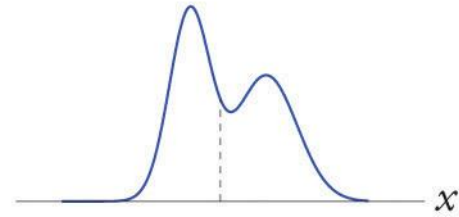
Sample Standard Deviation = $\frac{\text{Standard Deviation}}{n}$

OR

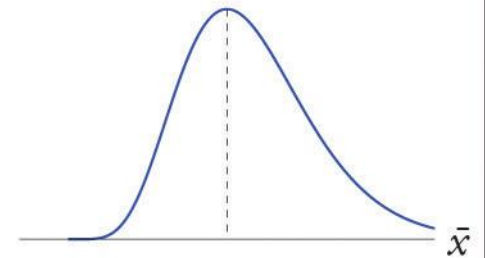
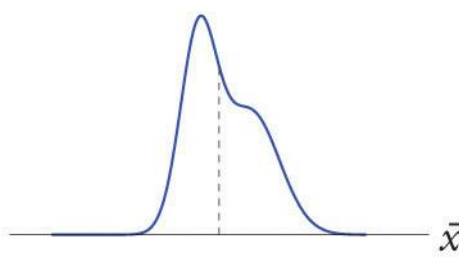
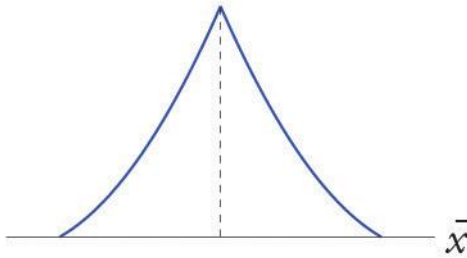
Sample Standard Deviation = $\frac{\sigma}{\sqrt{n}}$

Central Limit Theorem

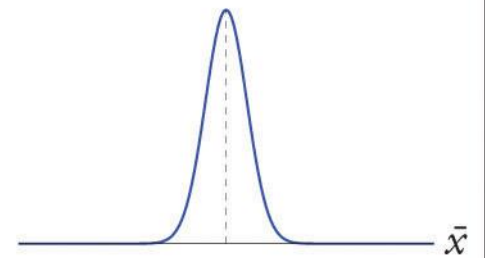
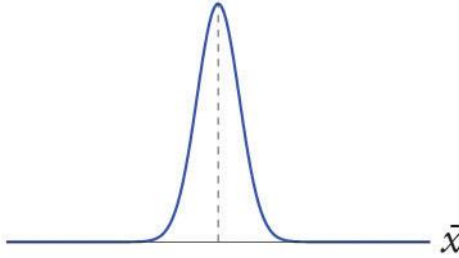
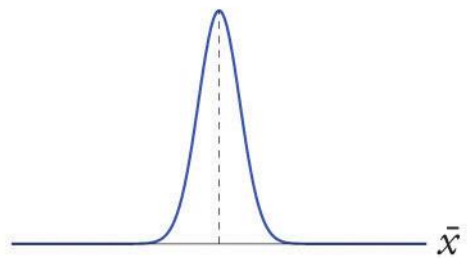
Population
distribution



Sampling
distribution
of \bar{X} with
 $n = 5$



Sampling
distribution
of \bar{X} with
 $n = 30$



Q1.

The mean and standard deviation of the tax value of all vehicles registered in a certain state are $\mu = \$13,525$ and $\sigma = \$4,180$. Suppose random samples of size 100 are drawn from the population of vehicles. What are the mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of the sample mean \bar{X} ?

Solution

Since $n = 100$, the formulas yield

$$\mu_{\bar{X}} = \mu = \$13,525$$

and

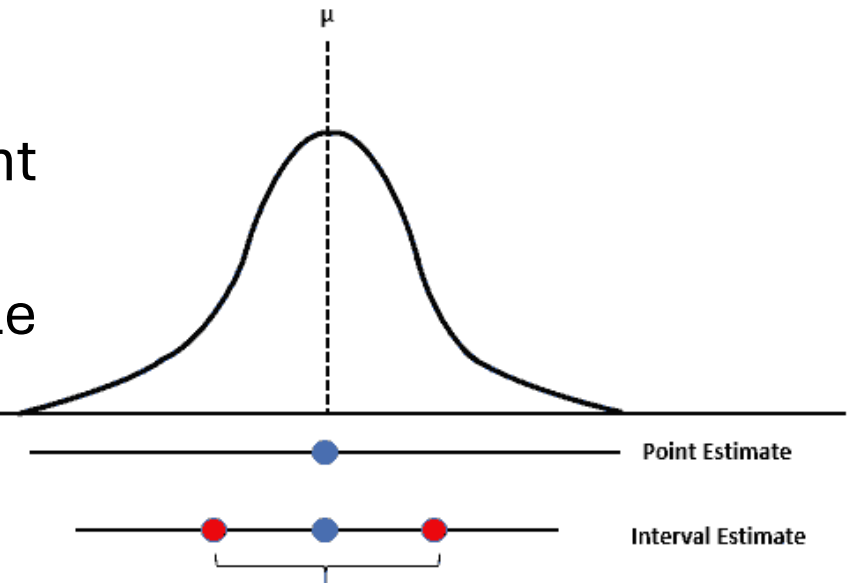
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4,180}{\sqrt{100}} = \$418$$

Point Estimate and Interval Estimate

- A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
- A point estimate is a sample statistic calculated using the sample data to estimate the most likely value of the corresponding unknown population parameter. In other words, we derive the point estimate from a single value in the sample and use it to estimate the population value.
- Take a sample, find \bar{x} . It is a close approximation of μ . But, depending on your sample size, that may not be a good point estimate.
- In fact, the probability that a single sample statistic is equal to the population parameter is very unlikely.

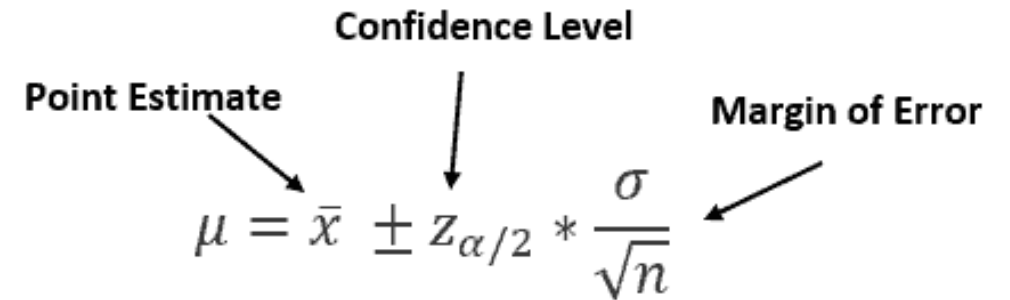
Point Estimate and Interval Estimate

- An interval estimate gives you a range of values where the parameter is expected to lie.
- A confidence interval estimate is a range of values constructed from sample data so that the population parameter will likely occur within the range at a specified probability. Accordingly, the specified probability is the level of confidence.
- Broader and probably more accurate than a point estimate
- Any parameter estimate that is based on a sample statistic has some amount of sampling error.



Point Estimate and Interval Estimate

A Confidence interval is used to express the precision and ambiguity of a particular sampling method.



The diagram shows the formula for a confidence interval: $\mu = \bar{x} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$. Three labels with arrows point to parts of the formula: 'Point Estimate' points to \bar{x} , 'Confidence Level' points to $z_{\alpha/2}$, and 'Margin of Error' points to the entire term $\pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$.

- A confidence *interval* is a range of values that probably contain the population mean.
- A Confidence level is a percentage of certainty that, in any given sample, that confidence interval will contain the population means.
- The Point estimate is a statistic (value from a sample) used to estimate a parameter (value from the population).
- The margin of error is the maximum expected difference between the actual population parameter and a sample estimate of the parameter. In other words, it is the range of values above and below sample statistics.