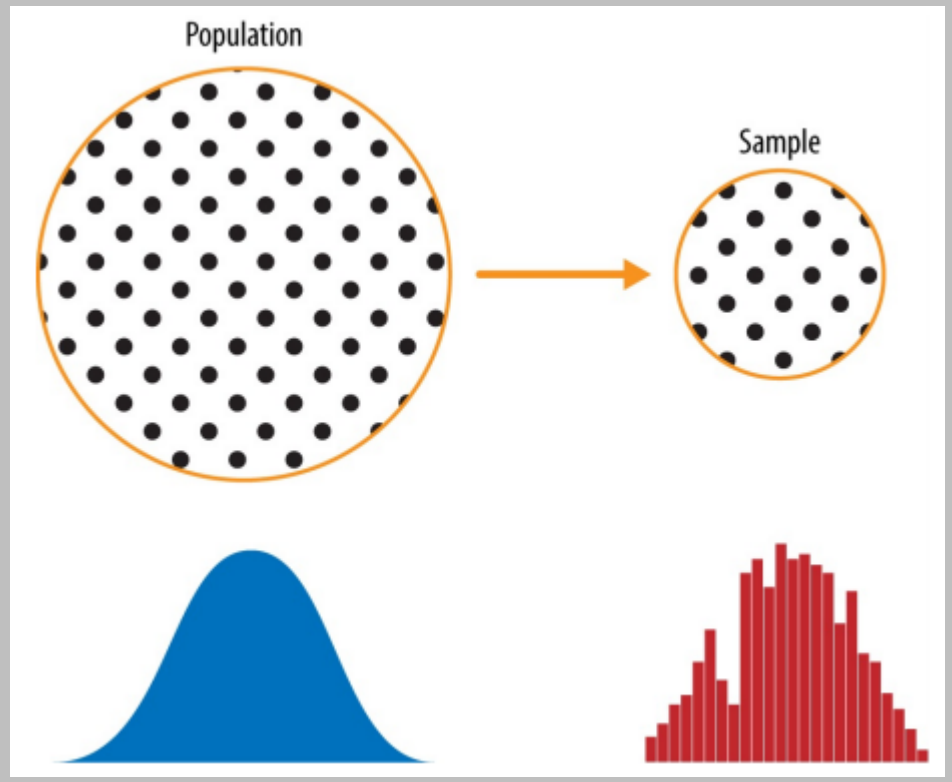
CSE303

Lecture 5: Sampling

SAMPLING

- Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.



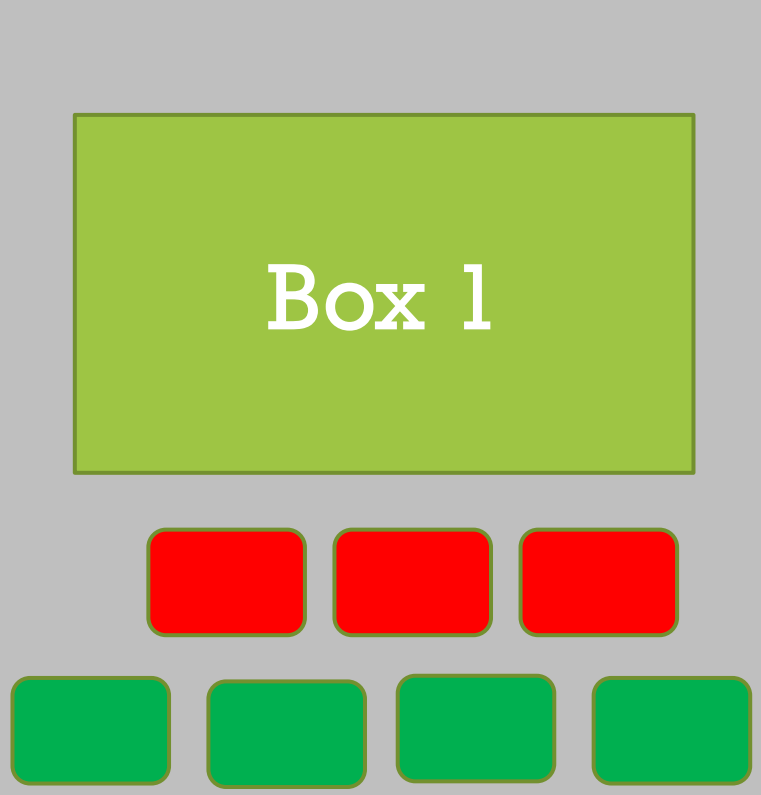
SOME KEY TERMS

- Sample: A subset from a larger data set.
- Population: The larger data set or idea of a data set.
- N: The size of the population. (number of elements in the whole dataset)
- n: The size of the sample. (number of elements in the subset)
- Random sampling: Drawing elements into a sample at random.
- Stratified sampling: Dividing the population into strata and randomly sampling from each strata.
- Simple random sample: The sample that results from random sampling without stratifying the population.
- Sample bias: A sample that misrepresents the population.

SIMPLE RANDOM SAMPLING

- Random sampling is a process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw.
- The sample that results is called a simple random sample.
- Sampling can be done with replacement, in which observations are put back in the population after each draw for possible future reselection.
- Or it can be done without replacement, in which case observations, once selected, are unavailable for future draws.

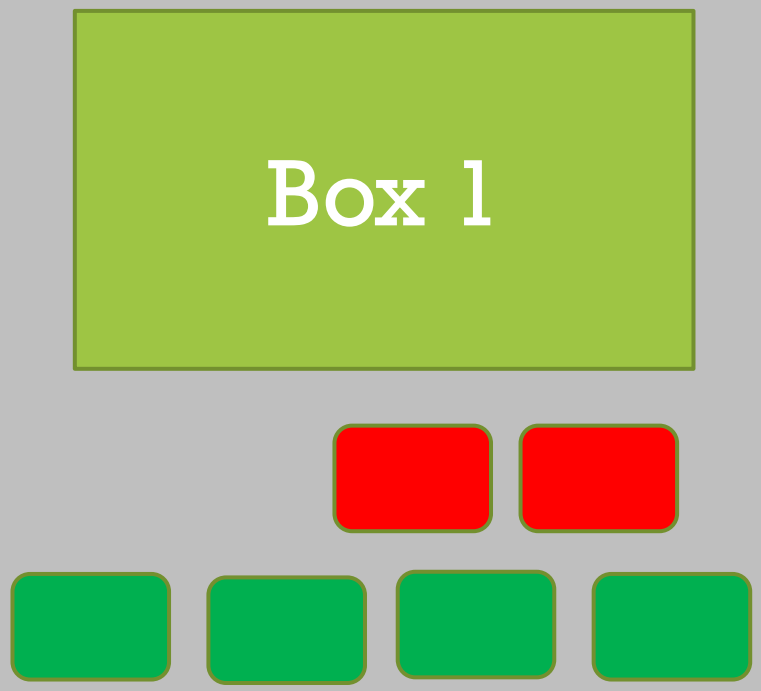
EXAMPLE: RANDOM SAMPLING WITH REPLACEMENT



Find the probability of picking two red toys from Box1 when you draw the samples with replacement.

P (first toy is RED) * P (second toy is RED) = 3/7 * 3/7 = 9/49

EXAMPLE: RANDOM SAMPLING WITHOUT REPLACEMENT

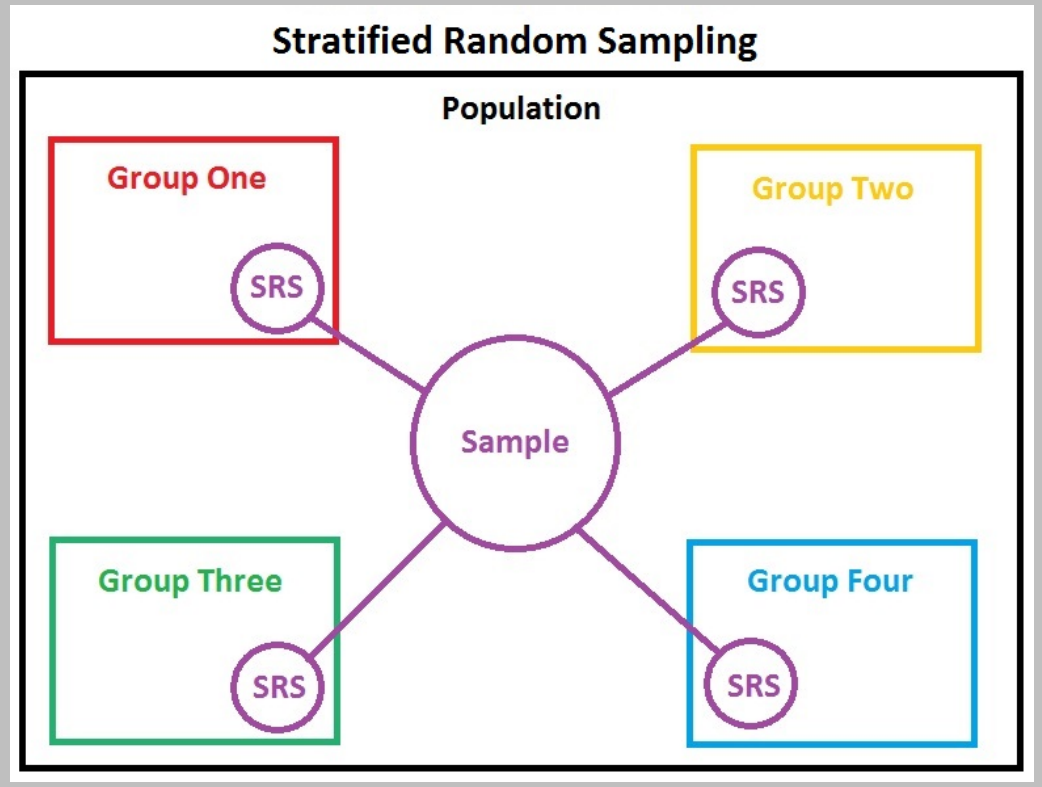


Find the probability of picking two red toys from Box1 when you draw the samples without replacement.

P (first toy is RED) * P (second toy is RED)
= 3/7 * 2/6= 6/42 = 1/7

STRATIFIED SAMPLING

- Stratification is the process of dividing members of the population into homogeneous subgroups before sampling. The strata should define a partition of the population.
- That is, it should be collectively exhaustive and mutually exclusive: every element in the population must be assigned to one and only one stratum/partition.

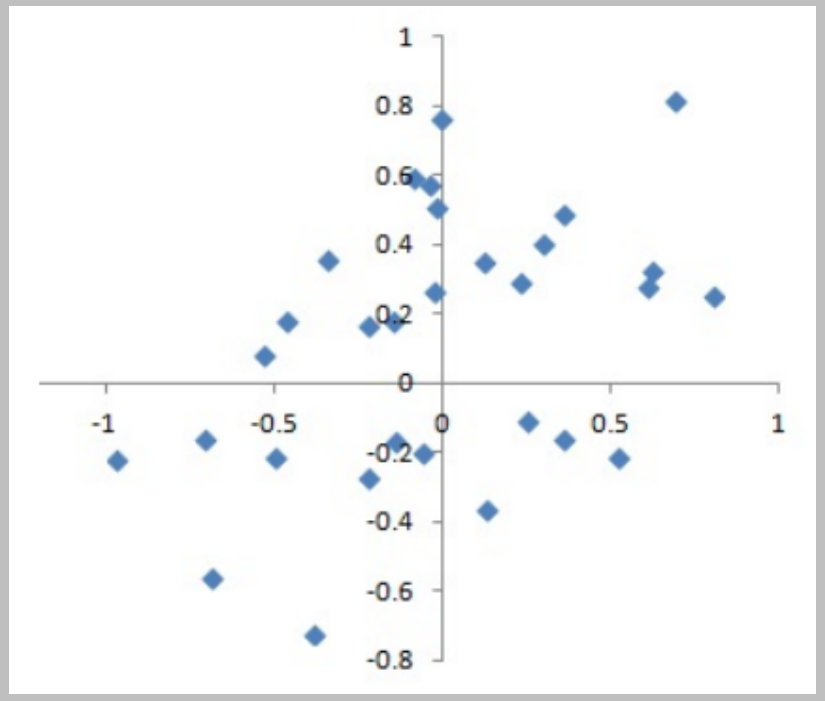


PROPERTIES OF STRATA/PARTITION IN STRATIFIED SAMPLING

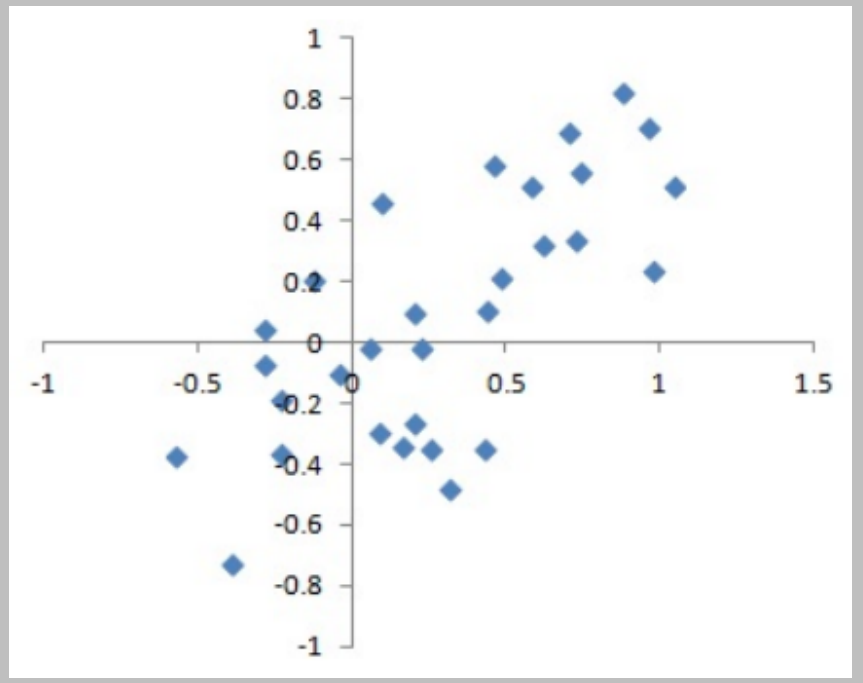
- Collectively Exhaustive: P1 U P2 U P3 = Entire Population
- Mutually Exclusive: P1 INTERSECT P2 INTERSECT P3 = NULL

BIAS

- Statistical bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process.



Gun shots with True Aim



Gun shots with Biased Aim

SAMPLE MEAN VS. POPULATION MEAN

- $\bar{x} = \frac{\sum x}{n}$ (Sample Mean) , Sample variance = $\frac{\sum(xi - \bar{x})^2}{n-1}$
- $\mu = \frac{\sum x}{N}$ (Population Mean), Population variance, sigma^2 = $\frac{\sum(xi - \bar{x})^2}{N}$

SAMPLE SIZE ESTIMATION

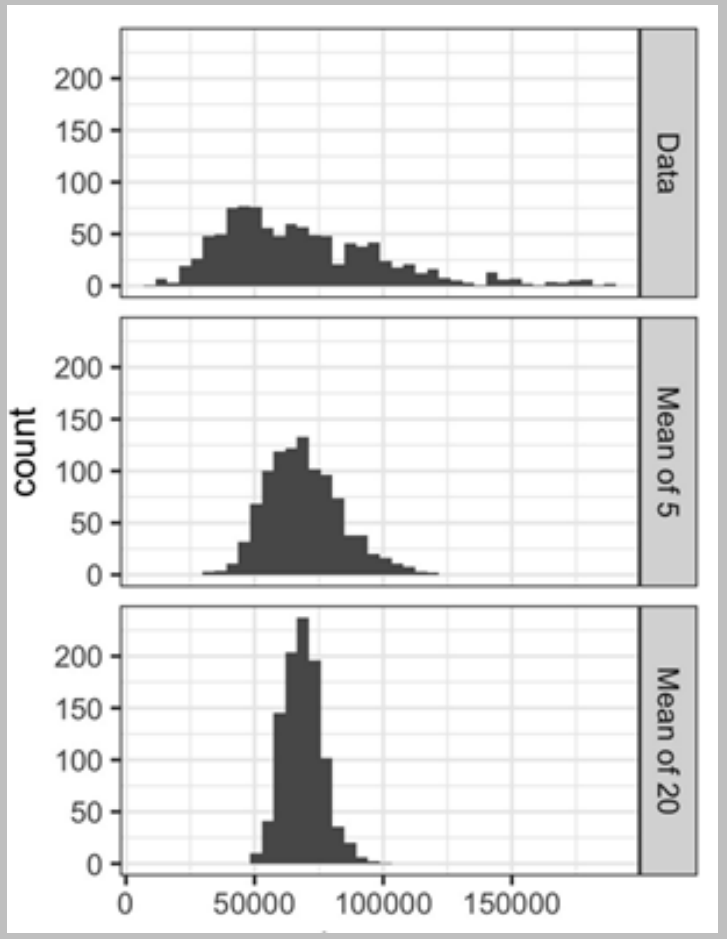
- The formula for calculating the sample size to estimate a population mean is given by:
- $n = (Z * \sigma / E)^2$
- Where:
  - n is the required sample size
  - Z is the Z-score corresponding to the desired confidence level (e.g., for a 95% confidence level, $Z \approx 1.96$)
  - σ is the estimated standard deviation of the population
  - E is the desired margin of error (the maximum difference between the sample mean and the population mean)

SAMPLE SIZE ESTIMATION

- The formula for calculating the sample size to estimate a population proportion is given by:
- $n = (Z^2 * p * (1 - p)) / E^2$
- Where:
  - n is the required sample size
  - Z is the Z-score corresponding to the desired confidence level
  - p is the estimated proportion of the population
  - E is the desired margin of error

SAMPLING DISTRIBUTION OF A STATISTICS

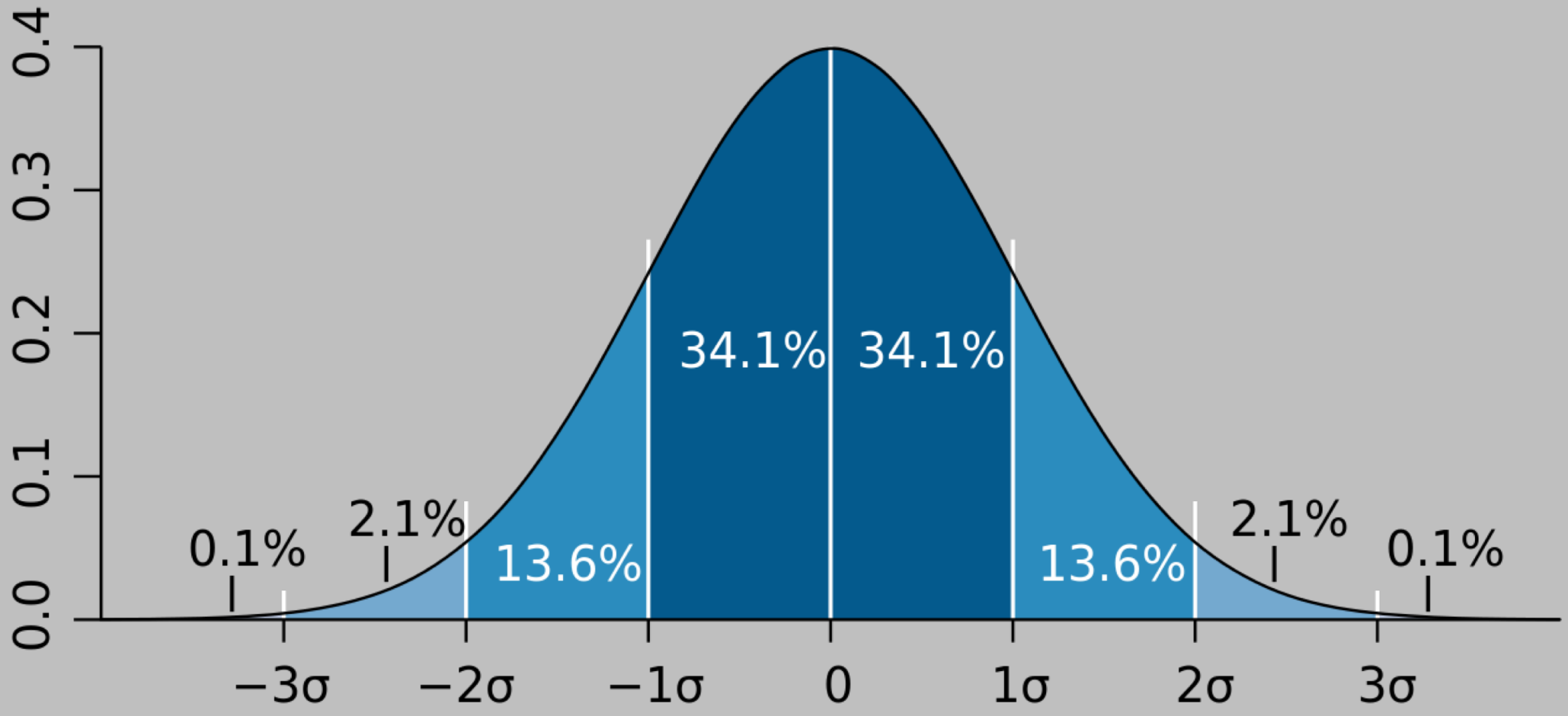
- The term sampling distribution of a statistic refers to the distribution of some sample statistic, overmany samples drawn from the same population.
- Sampling distribution of a sample mean:
  - Given a dataset, take a sample, and then compute the mean.
  - Repeat this step many times (10000)
  - Plot the sampling distribution
  - It will look like a bell-shaped curve



EXAMPLE

- Dataset – 10, 8, 3, 7, 1, 2, 5, 6, 9, 4
- Sample Size = 4
- Sample 1 (with replacement) → 2, 2, 4, 6 = 3.5
- Sample 2 (with replacement) → 9, 9, 10, 8 = 9.0
- …….
- Sample 10000 (with replacement) → 5, 6, 5, 7 = 5.75

NORMAL DISTRIBUTION



CENTRAL LIMIT THEOREM

- The tendency of the sampling distribution to take on a normal shape as sample size rises.
- We can prove it through an experiment!

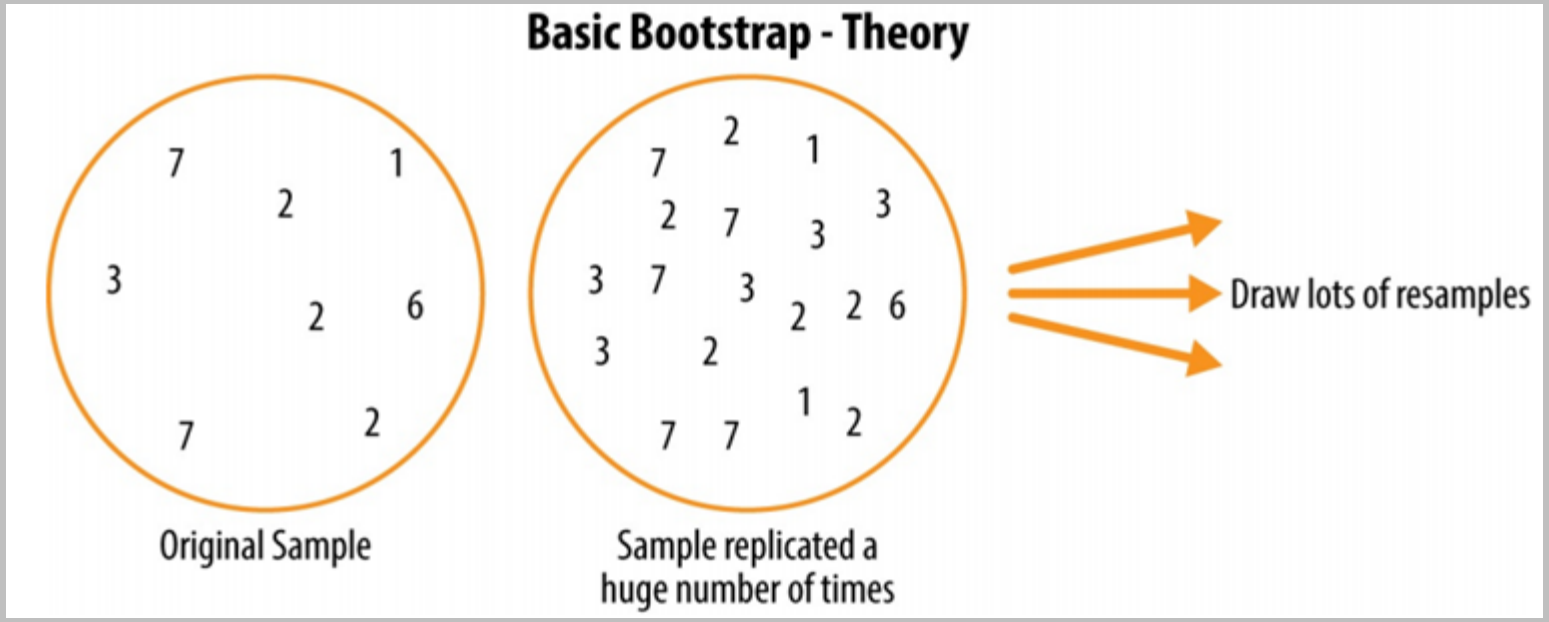
STANDARD ERROR

- The standard error is a single metric that sums up the variability in the sampling distribution for a statistic. The standard error can be estimated using a statistic based on the standard deviation s of the sample values, and the sample size n:

$$\text{Standard error} = SE = \frac{s}{\sqrt{n}}$$

SAMPLING DISTRIBUTION: THE BOOTSTRAP METHOD

- One easy and effective way to estimate the sampling distribution of a statistic, or of model parameters, is to draw additional samples, with replacement, from the sample itself and recalculate the statistic or model for each resample.



CALCULATING STANDARD ERROR USING BOOTSTRAP METHOD

- Draw a sample value, record, replace it.
- Repeat n times.
- Record the mean of the n resampled values.
- Repeat steps 1–3 for R times.
- Use the R results to:
  - Calculate their standard deviation (this estimates sample mean standard error).
  - Produce a histogram or boxplot.
  - Find a confidence interval.

CONFIDENCE INTERVAL

- A confidence interval, in statistics, refers to the probability (confidence level) that a population parameter will fall between a set of values for a certain proportion of times.
- Confidence intervals measure the degree of uncertainty or certainty in a sampling method.
- It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to survey the entire population.
- Confidence interval at 95% confidence level: It means that should you repeat an experiment or survey over and over again, 95 percent of the time your results will match the results you get from a population (in other words, your statistics would be sound!)

CONFIDENCE INTERVAL USING BOOTSTRAP METHOD

- Draw a random sample of size n with replacement from the data (a resample).
- Record the statistic of interest for the resample.
- Repeat steps 1–2 for many (R) times.
- For an x% confidence interval, trim [(1 – [x/100]) / 2]% of the R resample results from either end of the distribution.
- The trim points are the endpoints of an x% bootstrap confidence interval.

USEFUL RESOURCES

- Chapter 2, Practical Statistics for Data Scientists by Bruce and Bruce

THANK YOU