Big Data, Statistical Inference, and Practical Significance

Sampling Error

Nonsampling Error

Big Data

Understanding What Big Data Is

Big Data and Sampling Error

Big Data and the Precision of Confidence Intervals

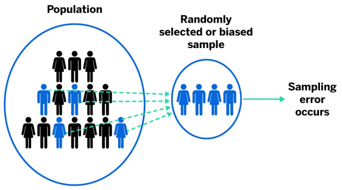
Implications of Big Data for Confidence Intervals

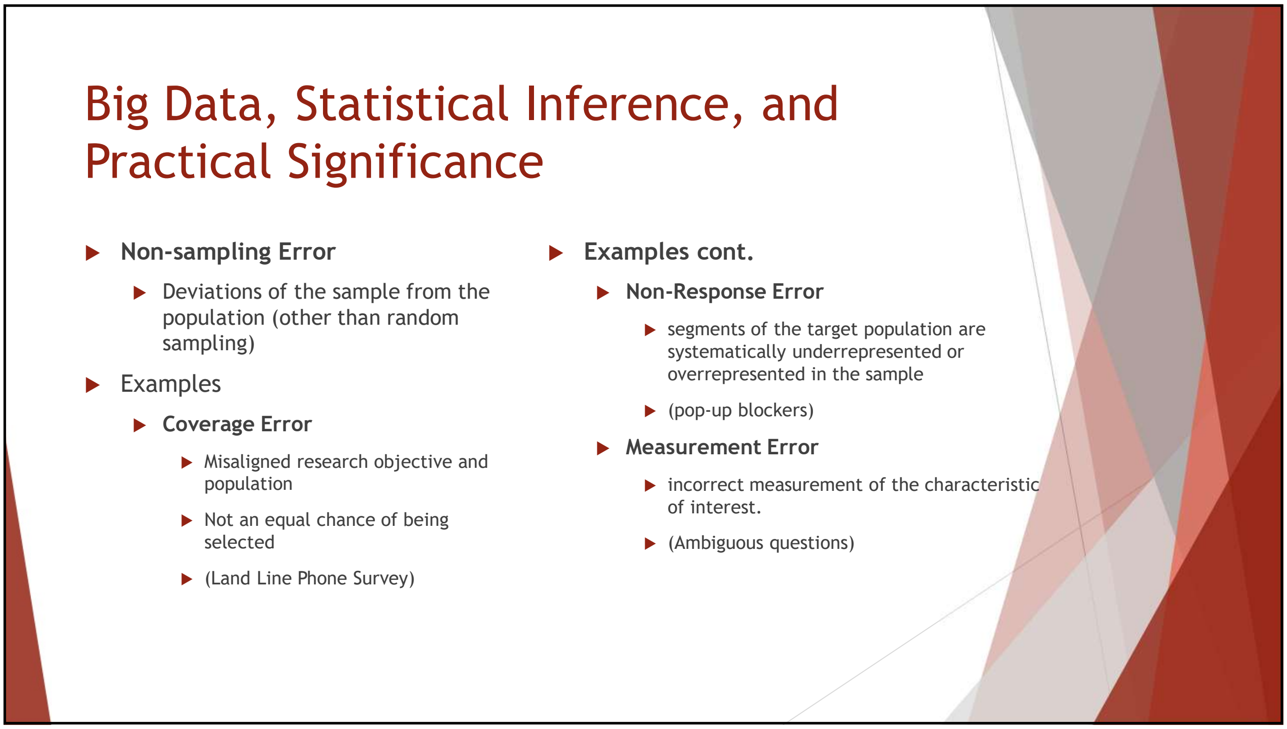
Big Data, Hypothesis Testing, and p Values

Implications of Big Data in Hypothesis Testing

Big Data, Statistical Inference, and Practical Significance

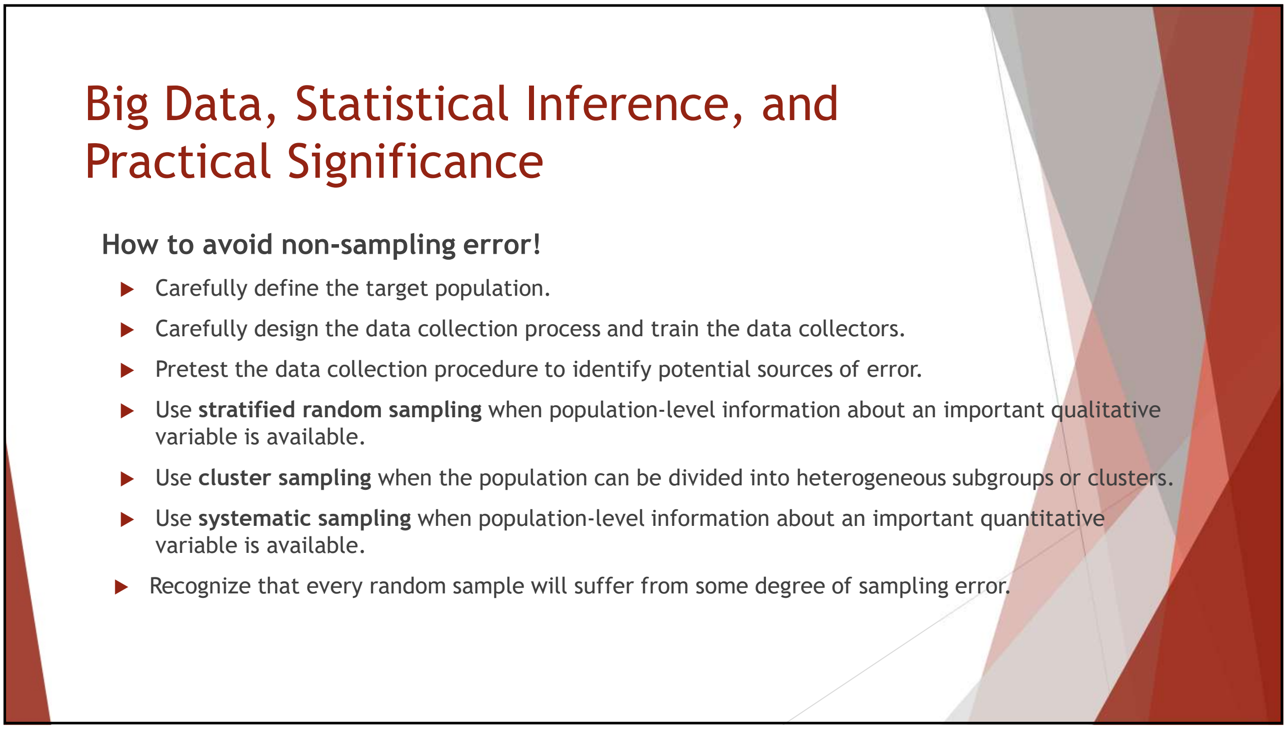
- **Sampling Error**
  - is the deviation of the sample from the population that results from random sampling.
- A random sample may not perfectly represent the population
- Sampling error is unavoidable when collecting a random sample.





Big Data, Statistical Inference, and Practical Significance

- **Non-sampling Error**
  - Deviations of the sample from the population (other than random sampling)
- Examples
  - **Coverage Error**
    - Misaligned research objective and population
    - Not an equal chance of being selected
    - (Land Line Phone Survey)
- **Examples cont.**
  - **Non-Response Error**
    - segments of the target population are systematically underrepresented or overrepresented in the sample
    - (pop-up blockers)
  - **Measurement Error**
    - incorrect measurement of the characteristic of interest.
    - (Ambiguous questions)



Big Data, Statistical Inference, and Practical Significance

How to avoid non-sampling error!

- Carefully define the target population.
- Carefully design the data collection process and train the data collectors.
- Pretest the data collection procedure to identify potential sources of error.
- Use **stratified random sampling** when population-level information about an important qualitative variable is available.
- Use **cluster sampling** when the population can be divided into heterogeneous subgroups or clusters.
- Use **systematic sampling** when population-level information about an important quantitative variable is available.
- Recognize that every random sample will suffer from some degree of sampling error.

Big Data, Statistical Inference, and Practical Significance

Big Data: Terminology for Describing the Size of Data Sets

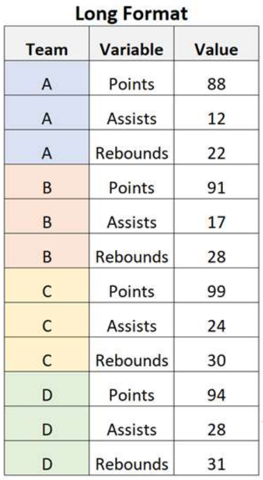
Number of Bytes	Metric	Name
1000^1	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

Big Data, Statistical Inference, and Practical Significance

Understanding What Big Data Is:

- Big data can be **tall (long) data**:

Warning: A data set can have so many observations that traditional statistical inference has little meaning.



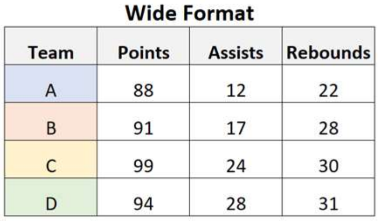
Long Format

Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

Big Data, Statistical Inference, and Practical Significance

Understanding What Big Data Is:

- Big data can also be **wide data:**



Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31

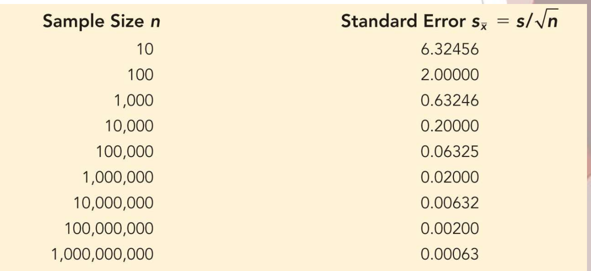
Warning: A data set that has so many variables that simultaneous consideration of all variables is infeasible.

Big Data, Statistical Inference, and Practical Significance

Implications of Big Data:

1. The standard error of the sampling distribution of the sample mean (proportion) **DECREASES** as the sample size **INCREASES.**

Implies: Very accurate estimate



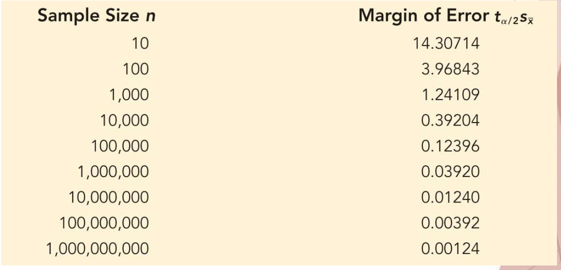
Sample Size n	Standard Error $s_{\bar{x}} = s/\sqrt{n}$
10	6.32456
100	2.00000
1,000	0.63246
10,000	0.20000
100,000	0.06325
1,000,000	0.02000
10,000,000	0.00632
100,000,000	0.00200
1,000,000,000	0.00063

Big Data, Statistical Inference, and Practical Significance

Implications of Big Data:

2. The sampling error **DECREASES** as sample size **INCREASE** also.

Implies: Very accurate estimate



Sample Size *n*	Margin of Error $t_{\alpha/2}s_{\bar{x}}$
10	14.30714
100	3.96843
1,000	1.24109
10,000	0.39204
100,000	0.12396
1,000,000	0.03920
10,000,000	0.01240
100,000,000	0.00392
1,000,000,000	0.00124

Big Data, Statistical Inference, and Practical Significance

Implications of Big Data for Confidence Intervals (cont.):

- Confidence intervals are extremely useful, but only effective when properly applied:
  - Interval estimates become increasingly precise as the sample size increases;
    - Extremely large samples will yield extremely precise estimates.
  - If there is non-sampling error
    - No interval estimate, no matter how precise, will accurately reflect the parameter being estimated
- When using interval estimation
  - Carefully consider whether a random sample of the population of interest has been taken.

Big Data, Hypothesis Testing, and *p* Values:

Problem!!

- *p* value also **DECREASES** as the sample size **INCREASES**.
- When sample size is large
  - Almost any difference between sample mean and hypothesized population mean will result in "Rejecting the Null"

Test of the Null Hypothesis $H_0: \mu \leq 84$ and Sample Mean $\bar{x} = 84.1$ Seconds

Sample Size *n*	t	*p* Value
10	0.01581	0.49386
100	0.05000	0.48011
1,000	0.15811	0.43720
10,000	0.50000	0.30854
100,000	1.58114	0.05692
1,000,000	5.00000	2.87E-07
10,000,000	15.81139	1.30E-56
100,000,000	50.00000	0.00E+00
1,000,000,000	158.11388	0.00E+00

Big Data, Statistical Inference, and Practical Significance

Implications of Big Data in Hypothesis Testing:

- The results of any hypothesis test, no matter the sample size,
  - are only reliable if the sample is relatively **free of nonsampling error.**
- Non-sampling error increases the likelihood of making a Type I or Type II error.
- When testing a hypothesis,
  - Ask yourself whether a random sample of the population of interest has been taken.
- No business decision should be based solely on statistical inference
  - Practical significance should always be considered in conjunction with statistical significance.