Big Data, Statistical Inference, and Practical Significance



Sampling Error Nonsampling Error Big Data

Understanding What Big Data Is Big Data and Sampling Error

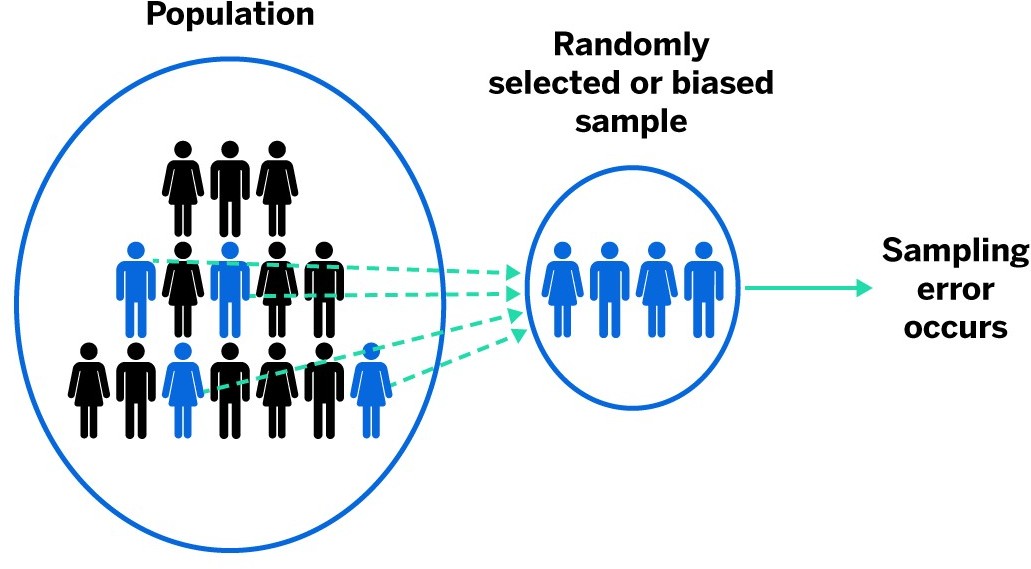
Big Data and the Precision of Confidence Intervals

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

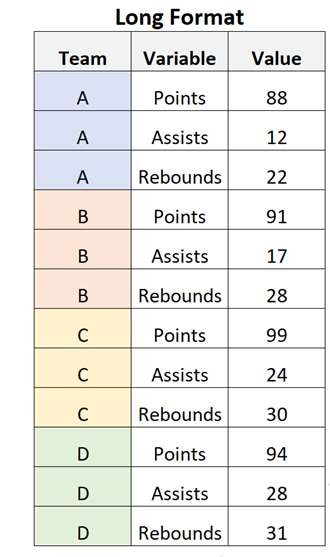
**How to avoid non-sampling error!**

* 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** |
| 10001 | kB | kilobyte |
| 10002 | MB | megabyte |
| 10003 | GB | gigabyte |
| 10004 | TB | terabyte |
| 10005 | PB | petabyte |
| 10006 | EB | exabyte |
| 10007 | ZB | zettabyte |
| 10008 | YB | yottabyte |

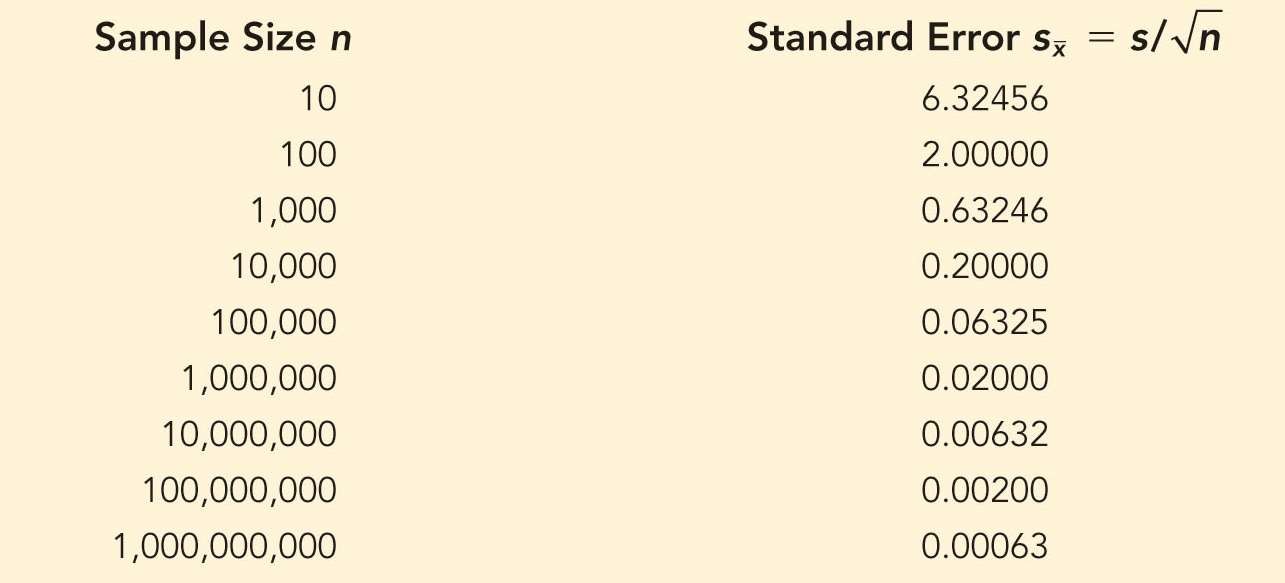
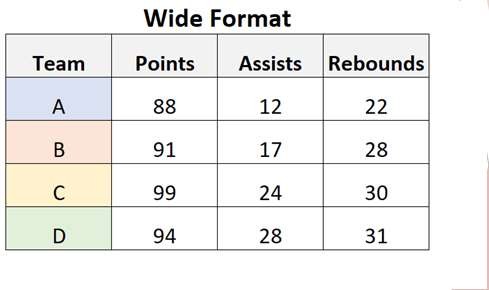


Big Data, Statistical Inference, and Practical Significance

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.



Big Data, Statistical Inference, and Practical Significance

Understanding What Big Data Is:

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

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

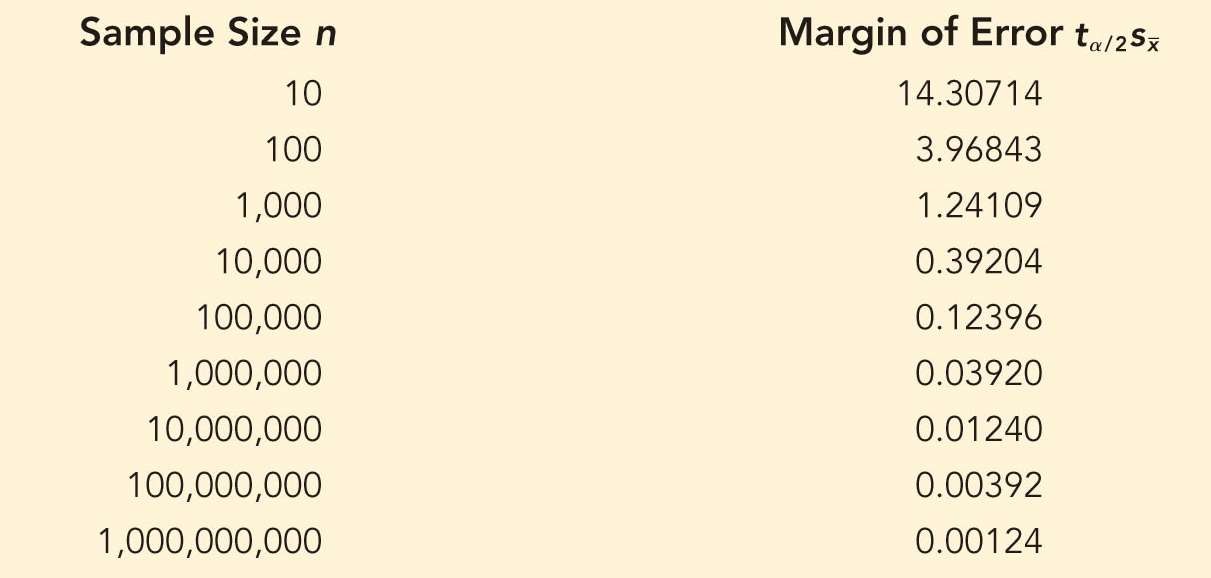


Big Data, Statistical Inference, and Practical Significance

**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**



Big Data, Statistical Inference, and Practical Significance

**Implications of Big Data:**

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

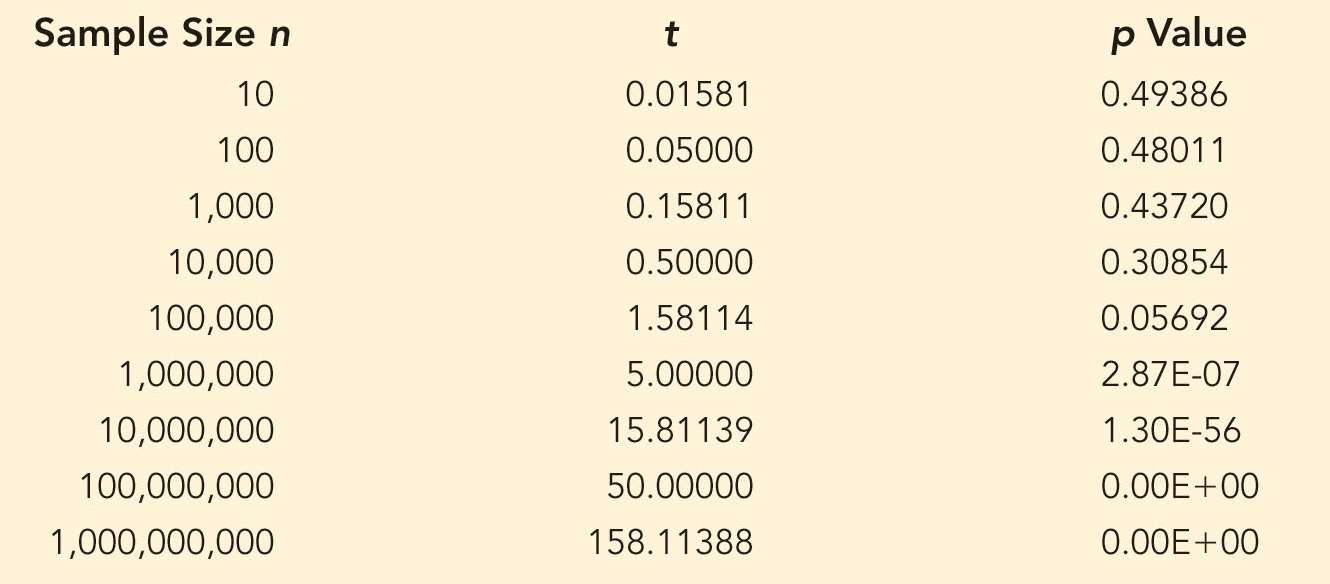
**Implies: Very accurate estimate**



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**.

Test of the Null Hypothesis 𝐻0: 𝜇 ≤ 84 and

Sample Mean 𝑥 = 84.1 Seconds

* When sample size is large
  + Almost any difference between sample mean and hypothesized population mean will result in “Rejecting the Null”



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.