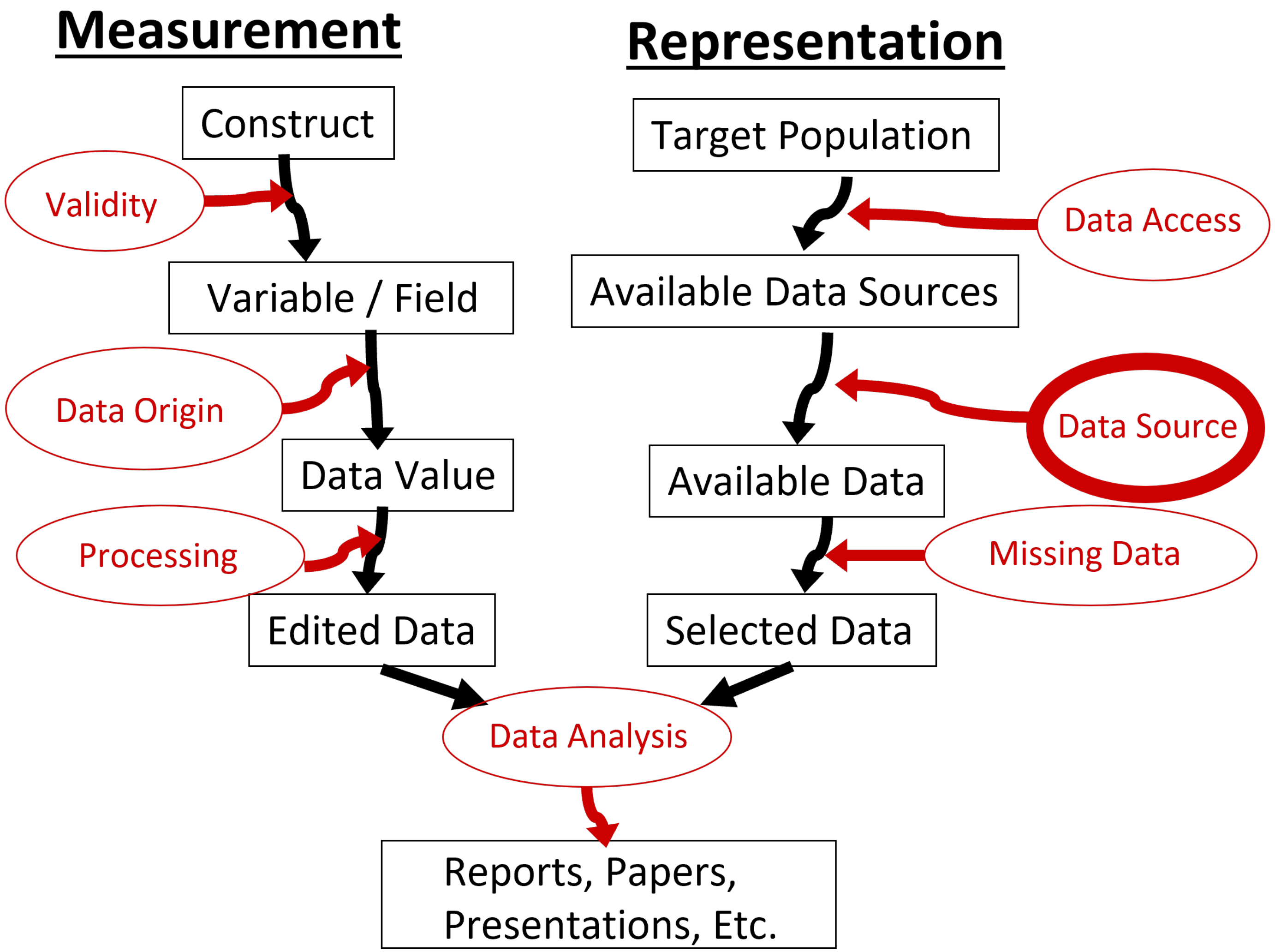
Data Source Threats to Designed Data

By James Wagner

Dimensions of TDQ

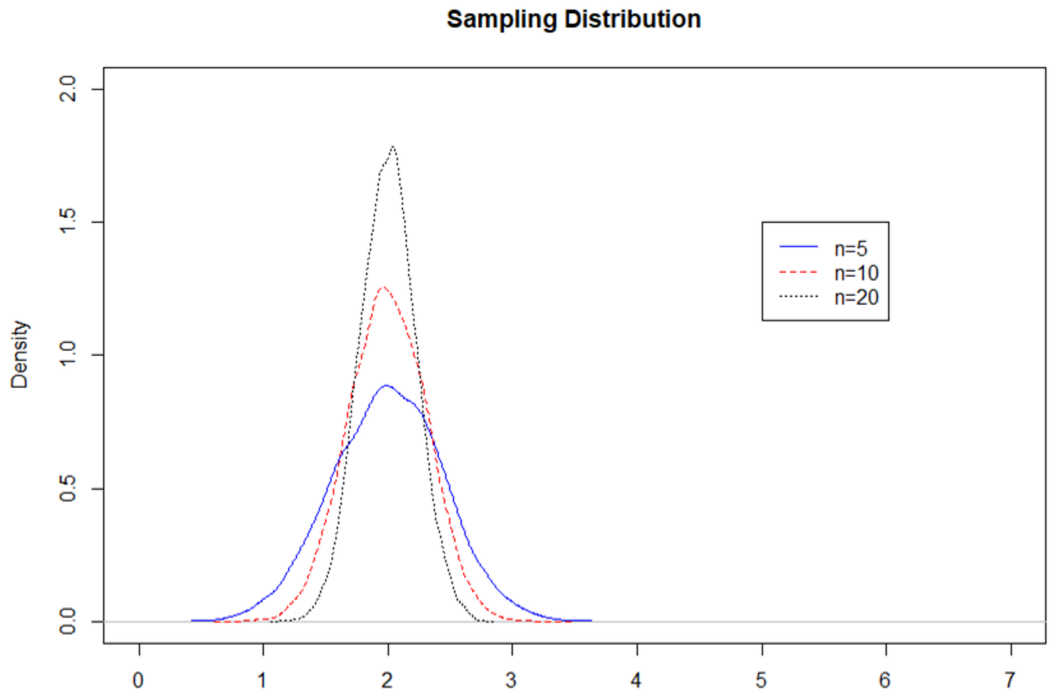


Data Source Threats to Designed Data

- For designed data, data source threats are usually related to **sampling** from a population.
  - As opposed to measuring the whole population.
- The error source is known as **sampling error.**

Data Source Threats to Designed Data (1)

- Hypothetical **sampling distribution** reflects the distribution of estimates that could be achieved under the same design.
- **Example:**
  - N=100,000, Population mean=2.
  - Drew 1,000 samples.
  - Estimate mean of each.
  - Plot the distribution of these 1,000 means.
  - Three different sample sizes:
    - n=5
    - n=10
    - n=20
- In general, larger sample size=smaller sampling error.



Data Source Threats to Designed Data (2)

- Sampling theory is a well-developed subfield within statistics.
- Several **sample design features** impact sampling error:
  - Clustering
  - Stratification
  - Weighting

Data Source Threats to Designed Data (3)

- **Clustering.**
- Often, cheaper to sample clusters.
- However, units within cluster may have **correlated measurements.**
- These correlations reduce the information in the sample relative to simple random sampling.

Data Source Threats to Designed Data (4)

- **Stratification.**
- **Strata** = groups on the sampling frame organized such that similar cases are in a stratum.
  - Need variables on sampling frame predictive of outcome variables.
- Stratification creates **efficiency.**
  - Conceptually, eliminates some possible samples, thereby narrowing the sampling distribution.
  - Example:
    - A population is 80% under the age of 65 and 20% 65+.
    - It is possible to randomly draw a sample that is *entirely* 65+.
      - Possible, but rare.
    - Stratification by age eliminates this possible sample.

Data Source Threats to Designed Data (5)

- **Weighting.**
- Complex sampling often produces **variable weights.**
- These weights may **increase** sampling error estimates.
- Example: Consider the following two designs.
  1. Draw a sample of n=100 from 300 million people in your country.
  2. Create two strata: your friends, everyone else.
     - You have 50 friends, and a sample of 50 more from the remaining 300 million people .
     - The two groups should get different weights (friends weight=1, other weights=(300,000,000-50)/50).
  - *Would you expect similar sampling error from these two designs?*

What’s next?

- Next, we will look at **data source threats** for gathered data.

UNIVERSITY OF MICHIGAN