Experimental optimization

Lecture 2: A/B testing I: Overview

David Sweet

Review

Expectation

- Game:
    - Flip a coin
    - Heads: You win $1
    - Tails: You lose $1
- What is the expectation?
- What is expectation?

Review

Expectation, sample mean

- RV is X
- *Expectation* of one play of the game:

$$E[X] = P\{\mathsf{H}\} \times \$1 + P\{\mathsf{T}\} \times (-\$1) = \$0$$

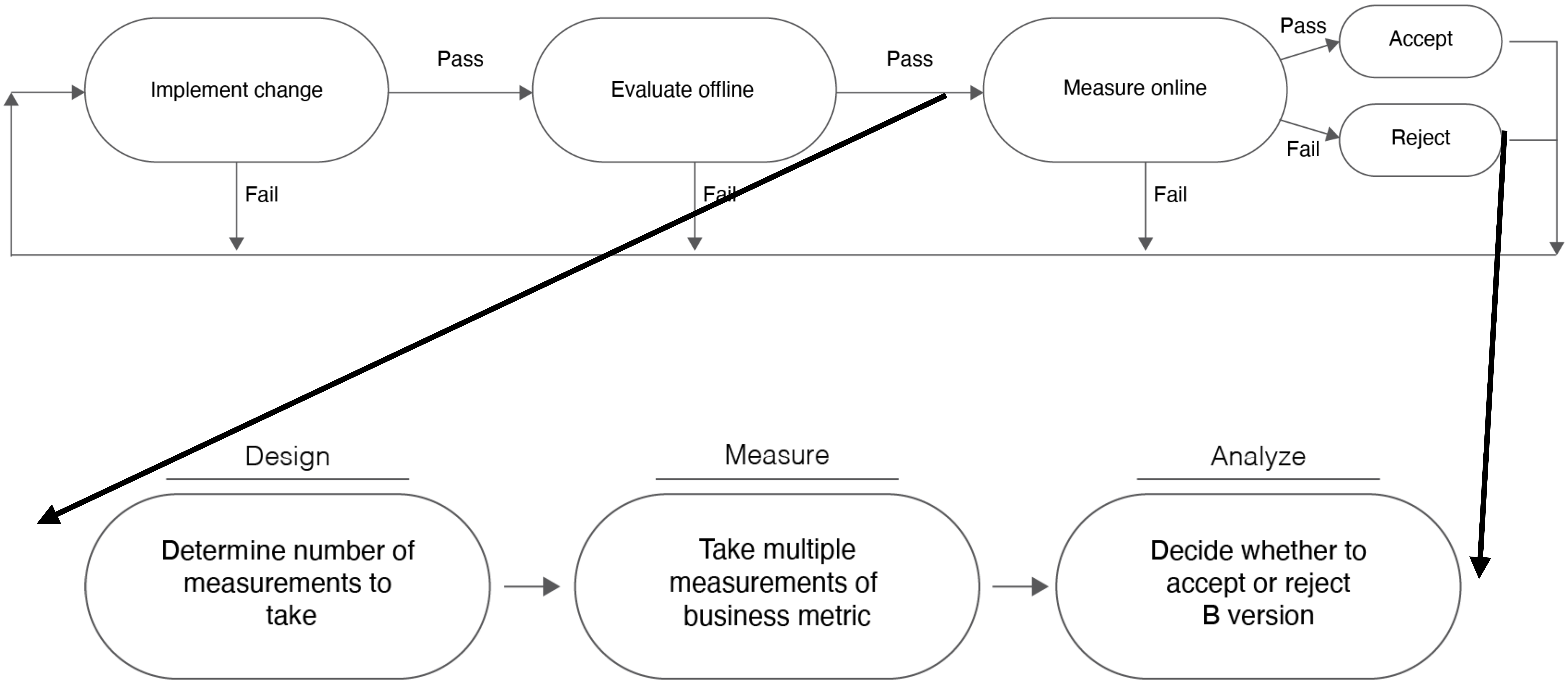
- Expectation is unobservable: One play of the game never returns $0.
- Estimate expectation by the sample mean over N plays:

$$E[X] \approx \Sigma_i x_i / N$$

What other unobservable quantities have you estimated?

Workflow / pipeline

Zoom in

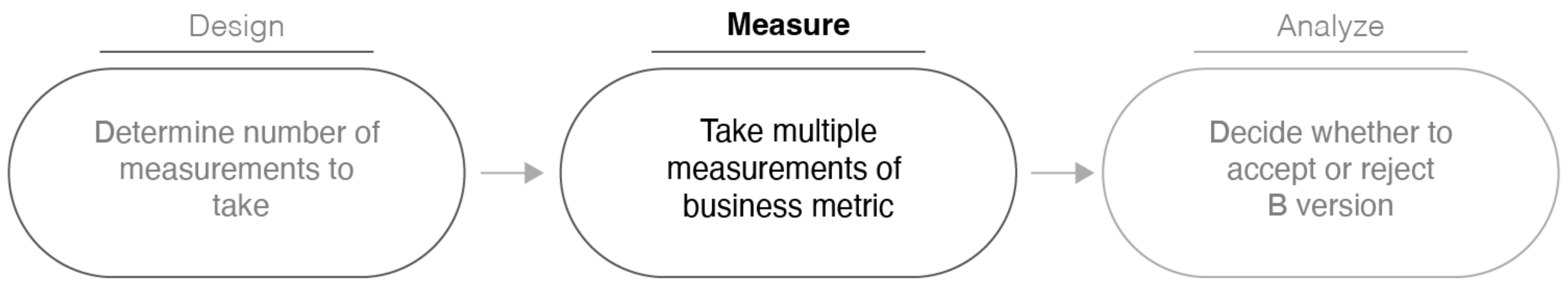


Aside: Simulators

- Book uses simulator as stand-in for real system
  - Python function s.t. `business_metric = f()`
- For class: Take real measurements on simulated system
- At work: Take real measurements on a real system
- Analogy:
  - In a class on regression, SL, NN, etc. you use sample data sets.
  - In (this) class on experimentation you'll use simulators.

Measure

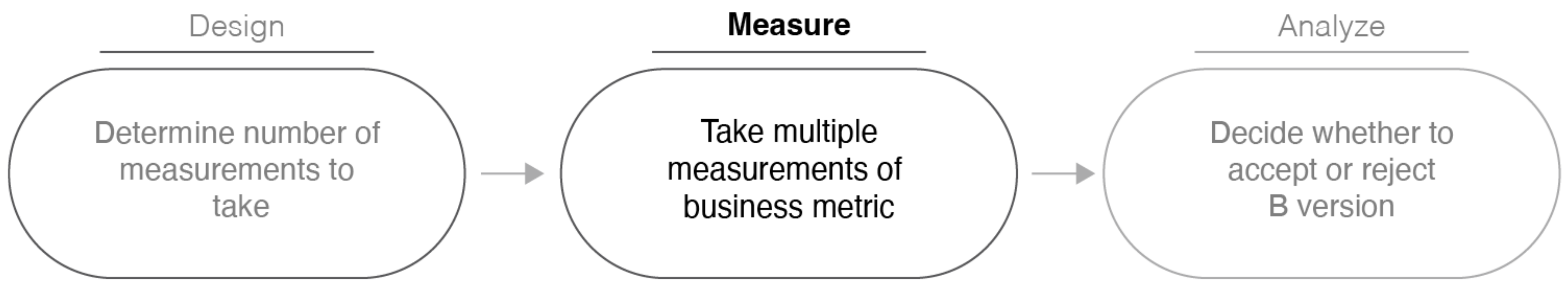
Record business metric values



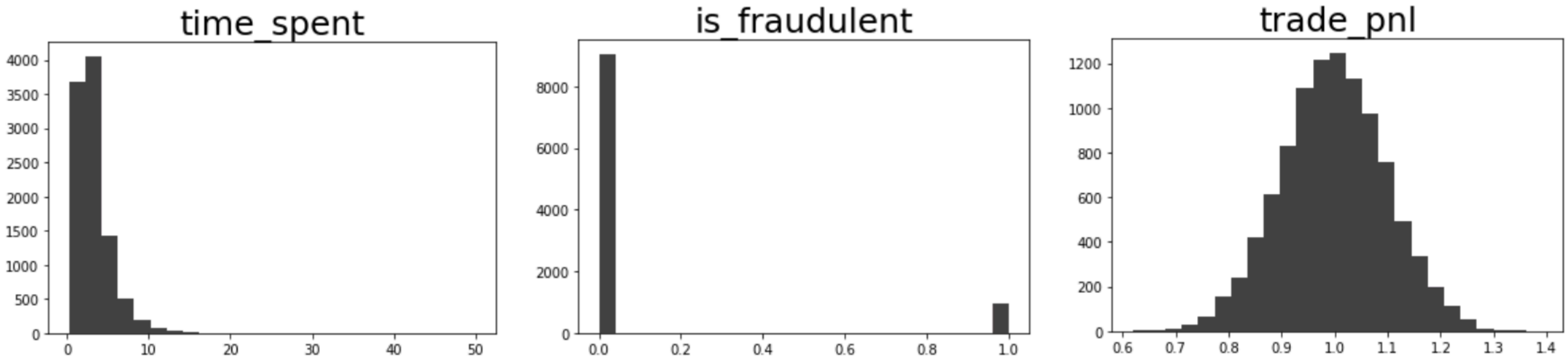
- Log values in production, post-process into BM

	Business metric	Values logged	Post process
Social media	Time spent per user per day	user id, date, time spent in a session	sum over sessions, avg. over users & dates
Credit card	P{fraud}	count of transactions, count of fraudulent	[num fraudulent] / [num transactions]
Trading strategy	PnL	trade prices and quantites	sum over returns on dollars held

Measure
Variation



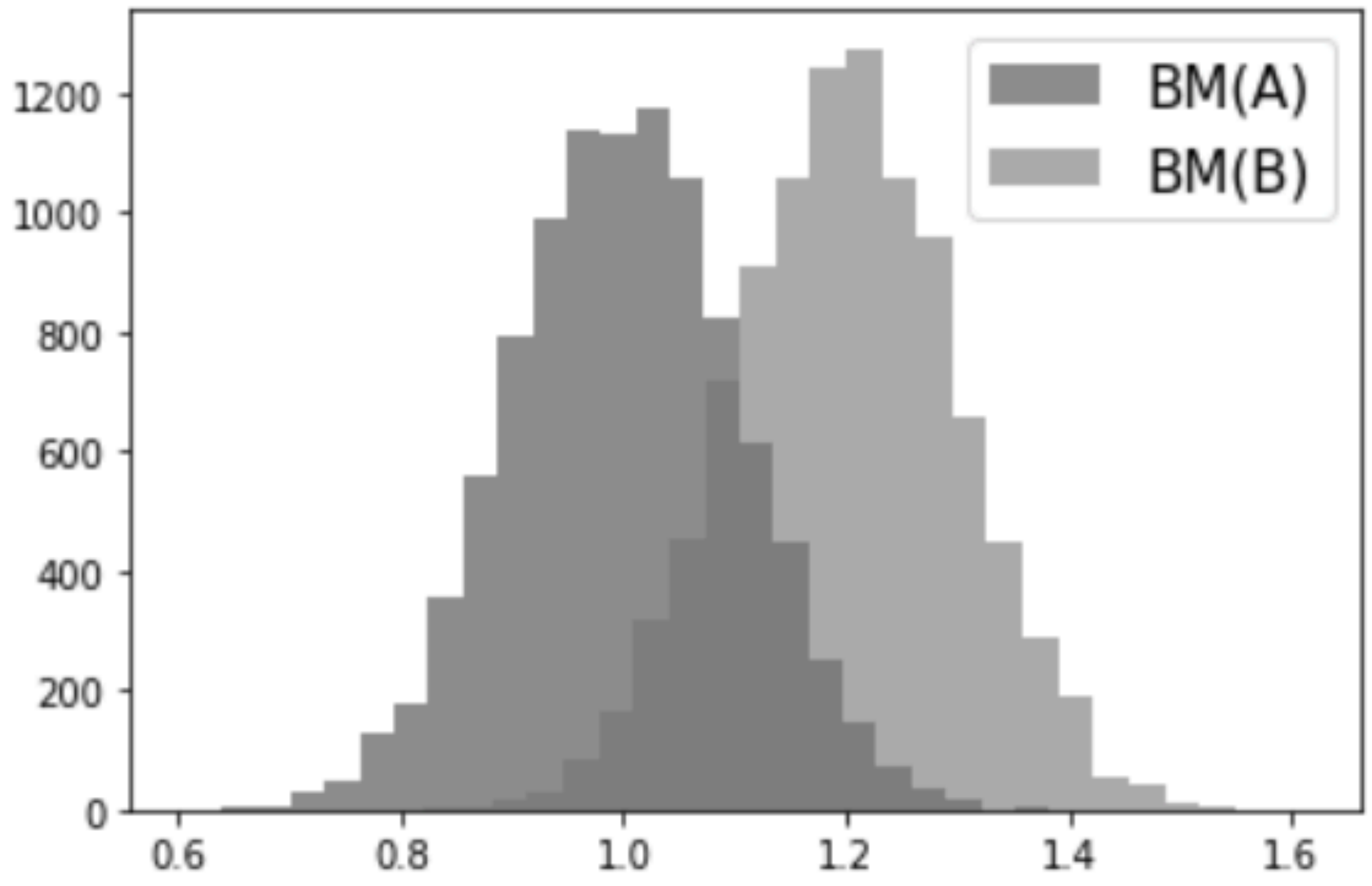
- Measured value varies from user to user, date to date, session to session, trade to trade, etc.



Analyze

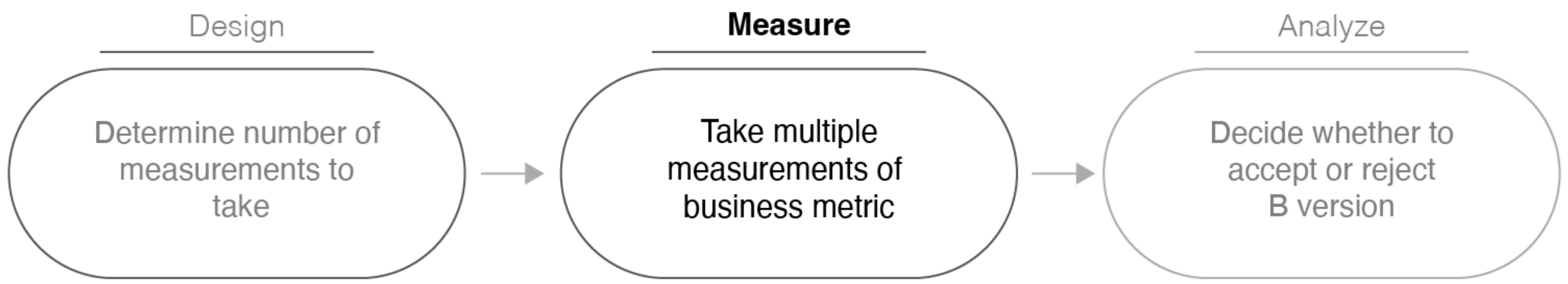
Compare measurements

- Measure business metric once for A and once for B
- Sometimes measure BM(A)>BM(B), sometimes BM(B)>BM(A)
  - ==> unreliable decisions



Measure II

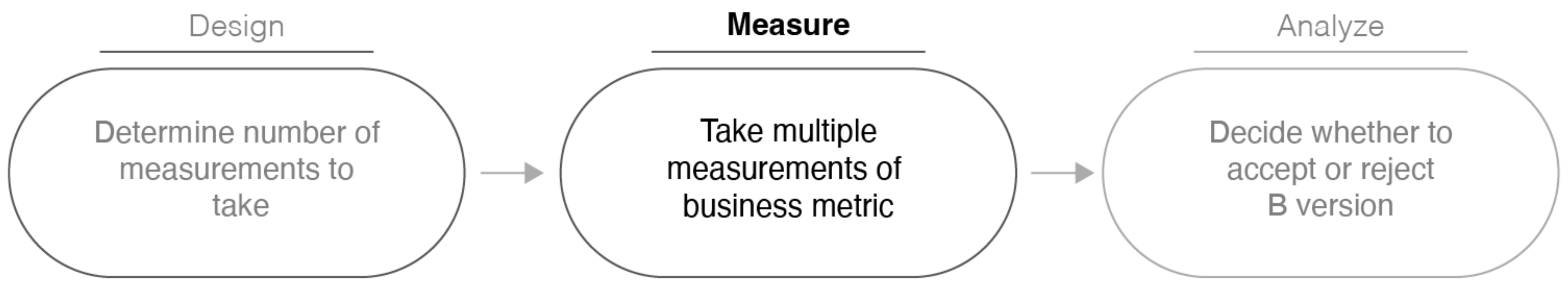
Replication reduces variation



- Replicate: Take multiple measurements and average them
- Measurements: $x_i \sim X, i = 1 \ldots N$
- Average: $\mu = \Sigma_i x_i / N$

Estimate the expectation of BM

- Replication reduces variation: $VAR(\mu) \leq VAR(X)$



Measure II
Standard error

- sample variance: $VAR(X) = \Sigma_i (x_i - \bar{x})^2 / N$
  estimate expectation: $\bar{x}$ by μ, $\hat{\bar{x}} = \mu$
  define: $\sigma^2 = VAR(X)$

- No “sample variance” of μ b/c we only have a single μ value

- Instead, estimate (Asn1,q2) $VAR(\mu)$ by $V\hat{A}R(\mu) = VAR(x)/N$

- Define standard error:

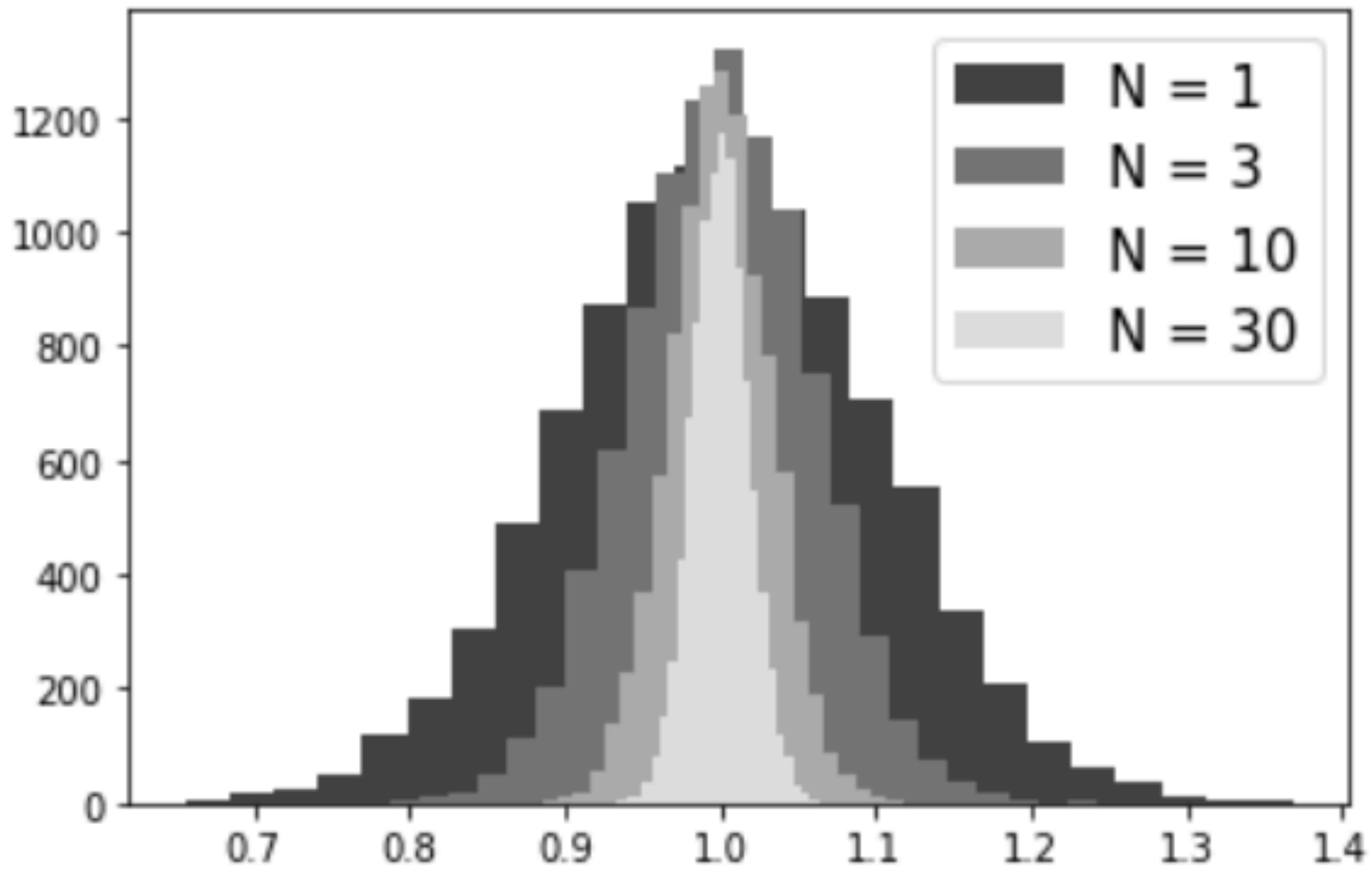
$$SE = \sqrt{VAR(\mu)}, \quad \hat{SE} = \sigma / \sqrt{N}$$

Measure II
Standard error

Design → Measure → Analyze

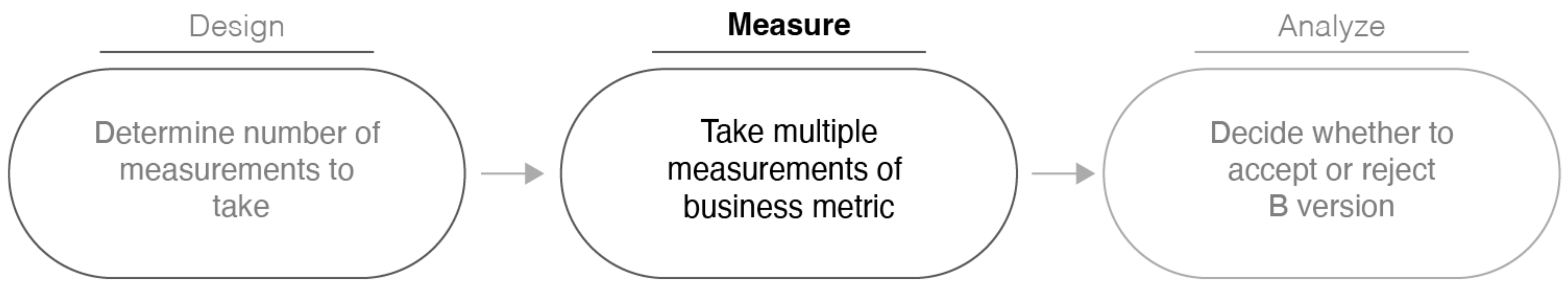
Determine number of measurements to take → **Take multiple measurements of business metric** → Decide whether to accept or reject B version

$$\hat{SE} = \sigma/\sqrt{N}$$



larger N == > smaller SE

Measure II
Standard error



- You can't control the variation in an individual measurement

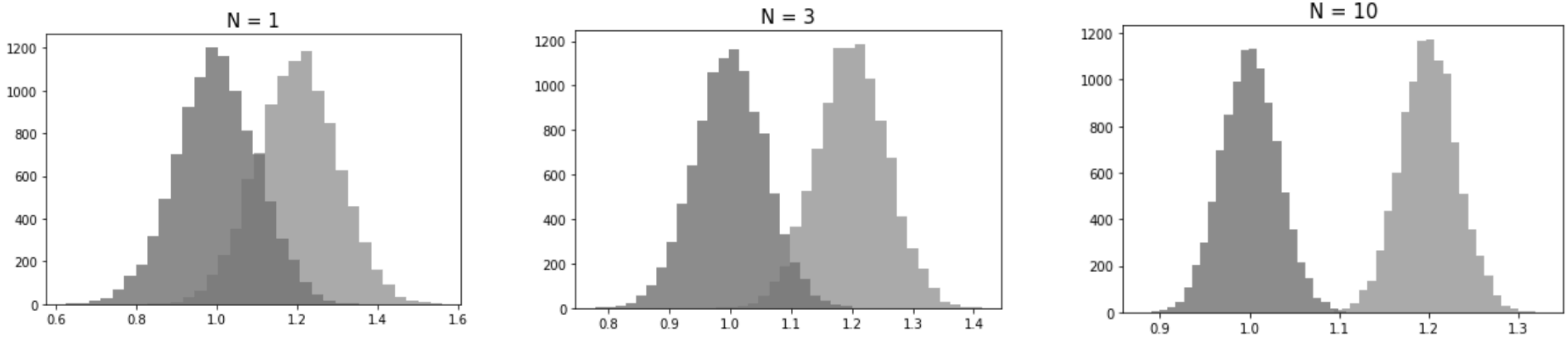
You can control the variation in an aggregate measurement.

- Setting N sets the level of variation, SE, in the aggregate measurement.
- *precision* ~ 1/SE

Analyze II

Compare aggregate measurements

- *individual measurement*: x_i N=1
- *aggregate measurement*: μ, N>1 ==> more reliable decisions

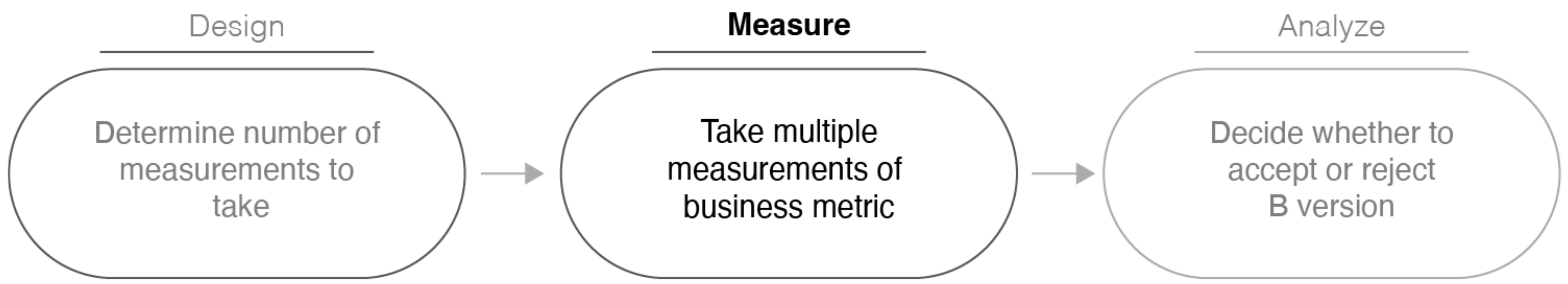


Design

Minimize experimentation costs

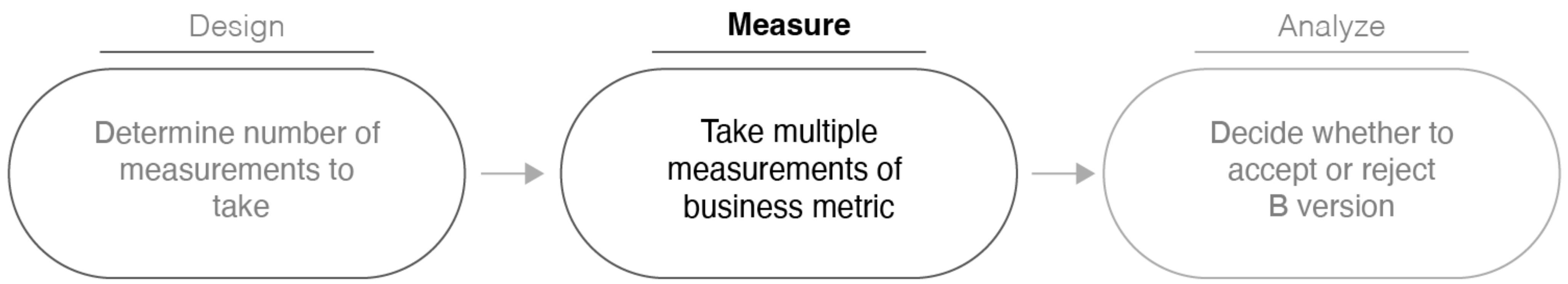
- Tradeoff:
  - Larger N gives lower SE
  - Smaller N gives lower experimentation costs
- A/B test design optimizes N
  - Smallest N s.t. SE is “small enough”
  - (“small enough” discussed next lecture)

Measure III
Bias



- Example, credit card fraud detection system, BM = P{fraud}
  - version A: old ML model
  - version B: new ML model
- A/B test: Collect large N of BM(A), BM(B)
- Configure to run A in US and B in Europe
- BM(B) < BM(A) by a lot!

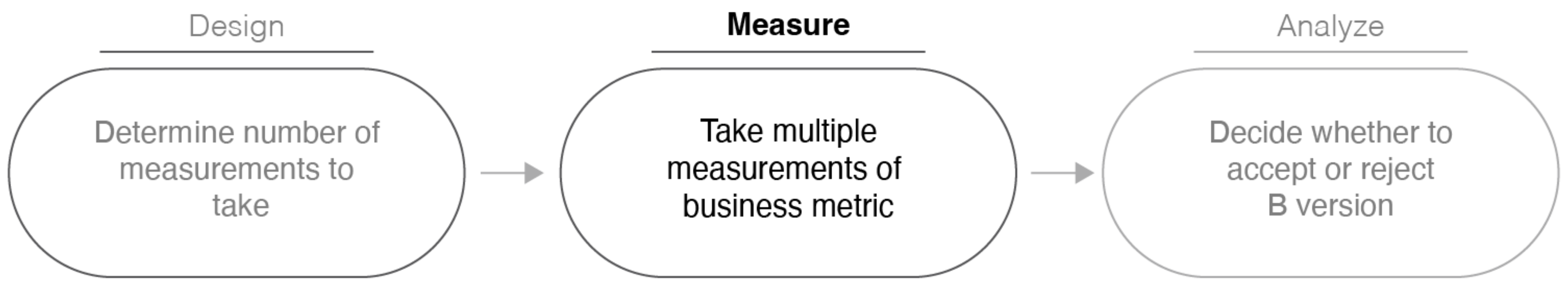
Measure III
Confounder bias



- Europe has EMV chip-card law.
- If you ran A in US and A in Europe, BM(A, Europe) < BM(A, US)
- So is B better than A, or is it just that Europe is better than US?
- Country (US/Europe) is a *confounder*
- Could fix by running A,B in both US and Europe.
- But what about the other confounders?

Do (could) we even know what they are?

Measure III



Randomization removes confounder bias

- Randomization:
  - Flip a coin every time a transaction enters the system.
  - Heads, use A
  - Tails, use B
- Randomization makes measurements *accurate*, i.e. unbiased
- Run for all transactions (US, Europe, etc.)

Don’t need to know what the confounders are!

A/B Testing

Summary

- Replication makes a measurement precise
- Randomization makes a measurement accurate
- A/B test design minimizes the experimentation cost for a measurement of a given precision
- An analogy:

Variation : precision : replication

:: Bias : accuracy : randomization