

# **The ABC of statistics**

Jonas Ranstam

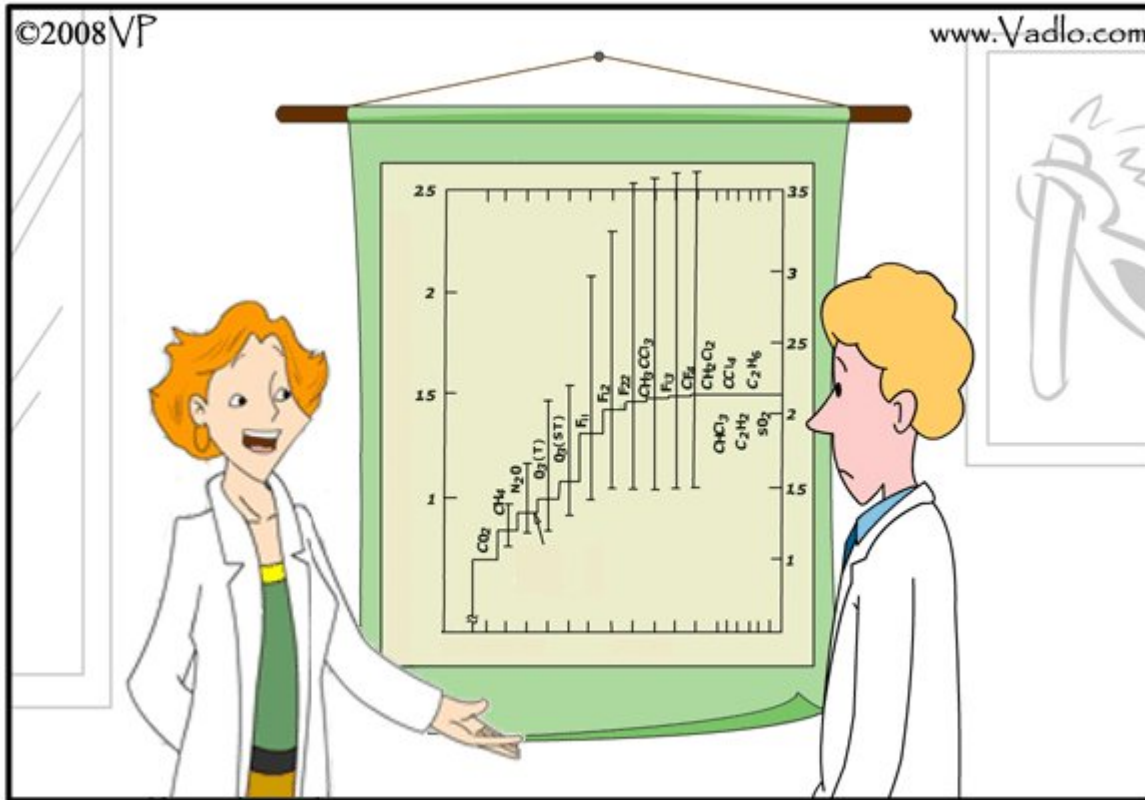
The epidemiologists  
**Have they got scares  
for you!**



John Brignell

©2008VP

www.Vadlo.com

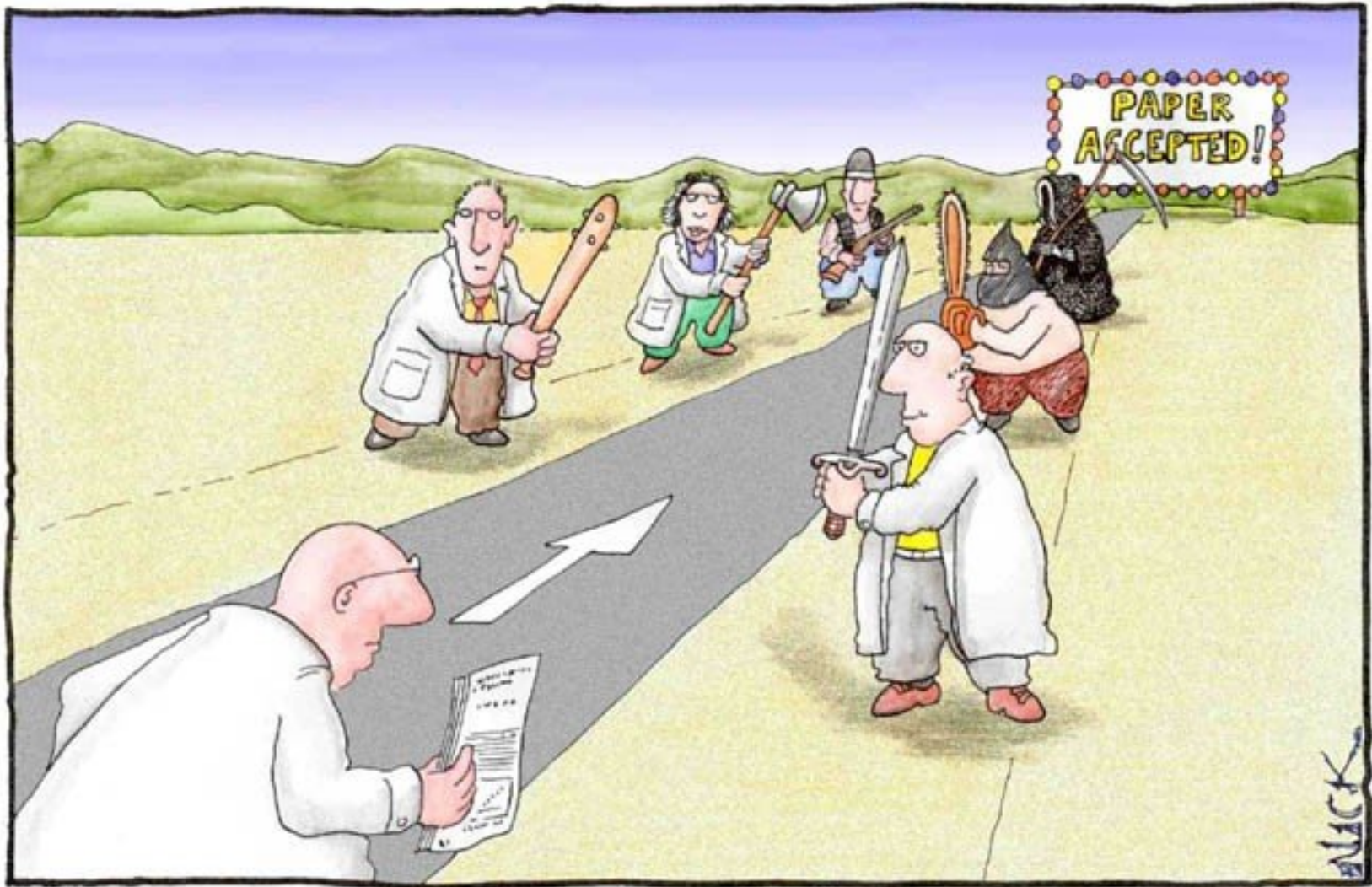


*Did you really have to show the error bars?*

# A scientific report

The idea is to try and give all the information to help others to judge the value of your contributions, not just the information that leads to judgment in one particular direction or another.

*Richard P. Feynman*

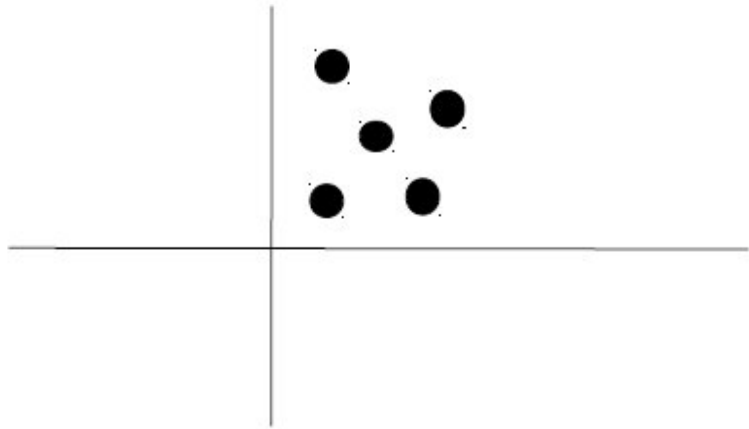


# Uncertainty is ubiquitous

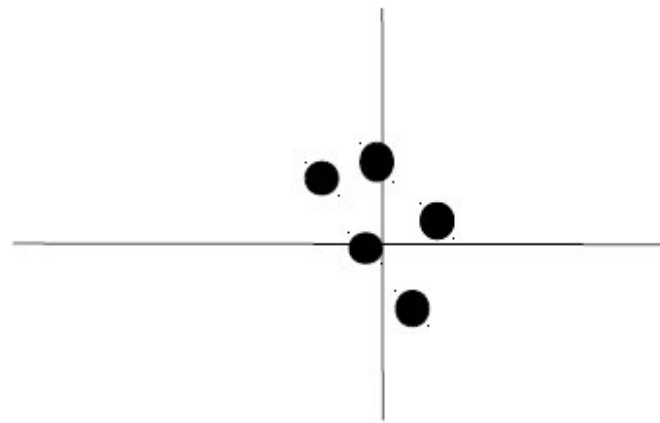
1. Random variation (precision)
  - a) measurement errors
  - b) sampling variation
  
2. Systematic deviation (validity)
  - a) selection bias
  - b) information bias
  - c) confounding bias

# Precision and validity of estimates

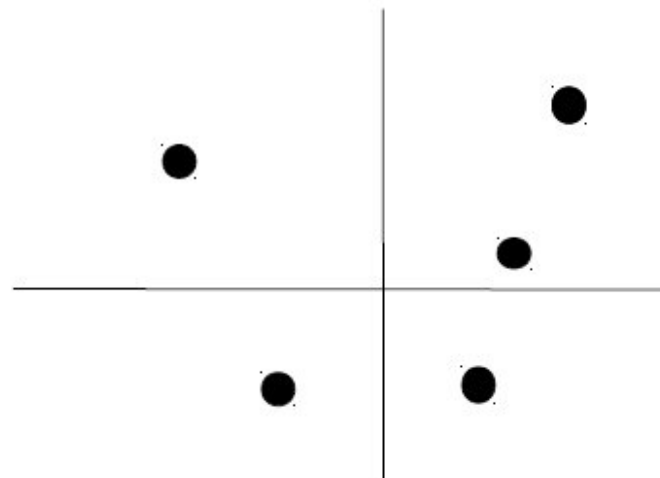
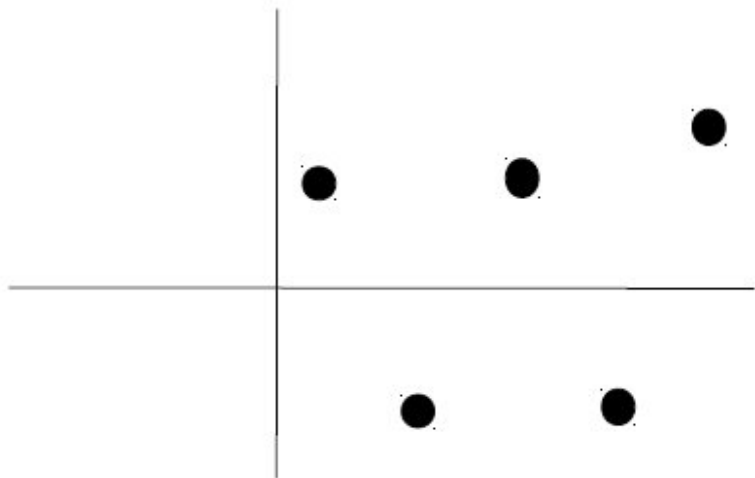
Lower validity



Higher validity



Higher precision



Lower precision

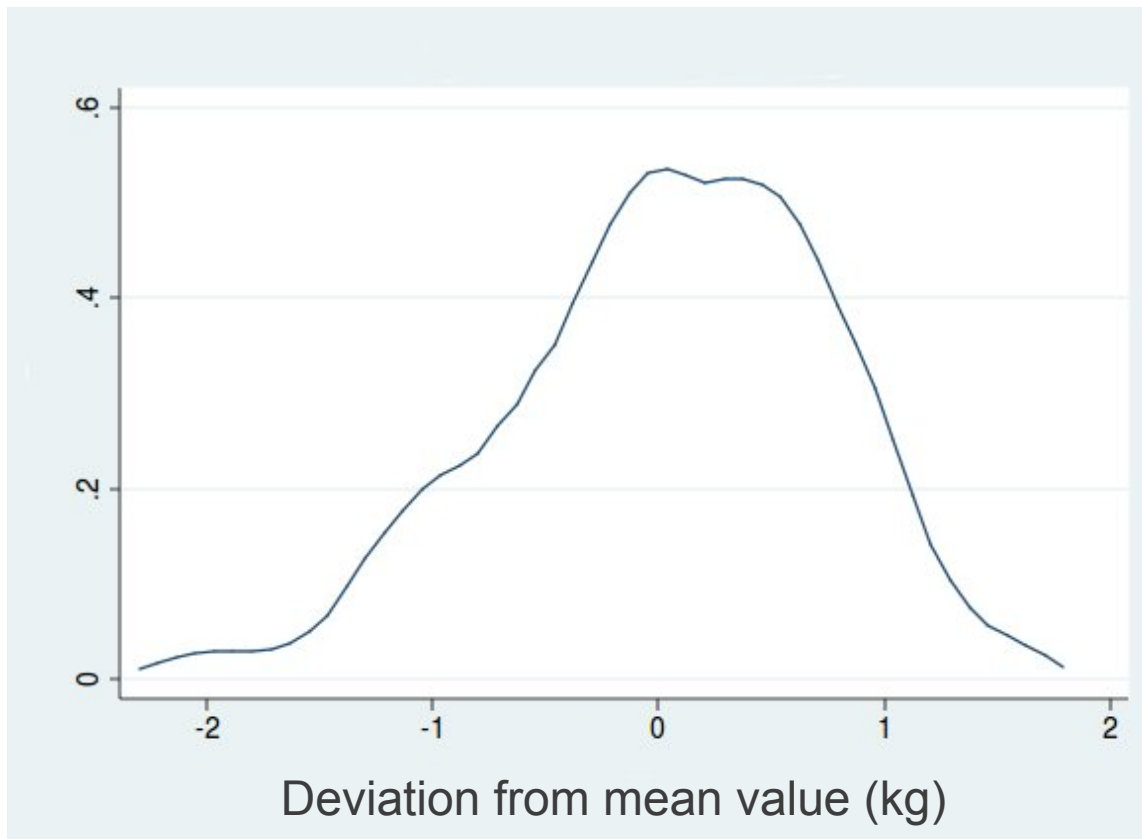
# Random variation

## Measurement errors

How uncertain is a body weight measurement?

# Error distribution

Variation in observed body weight during 133 consecutive daily measurements (residuals, detrended using lowess)



## How uncertain is an observed weight of 77kg?

---

### Degree of uncertainty

---

68.0%	$\pm 1.5\text{kg}$
95.0%	$\pm 3.0\text{kg}$
99.7%	$\pm 4.5\text{kg}$

---

# How uncertain is a weight change, between two consecutive measurements?

---

## Degree of uncertainty

---

68.0%	$\pm 2.1\text{kg}$
95.0%	$\pm 4.2\text{kg}$
99.7%	$\pm 6.4\text{kg}$

---

# Random variation


## Sampling variation

How uncertain is a mean value,

or, what does “statistically significant” mean?

# Statistics

Numerical descriptions



Observed  
sample



## Significance related statements

“There was no difference in...”

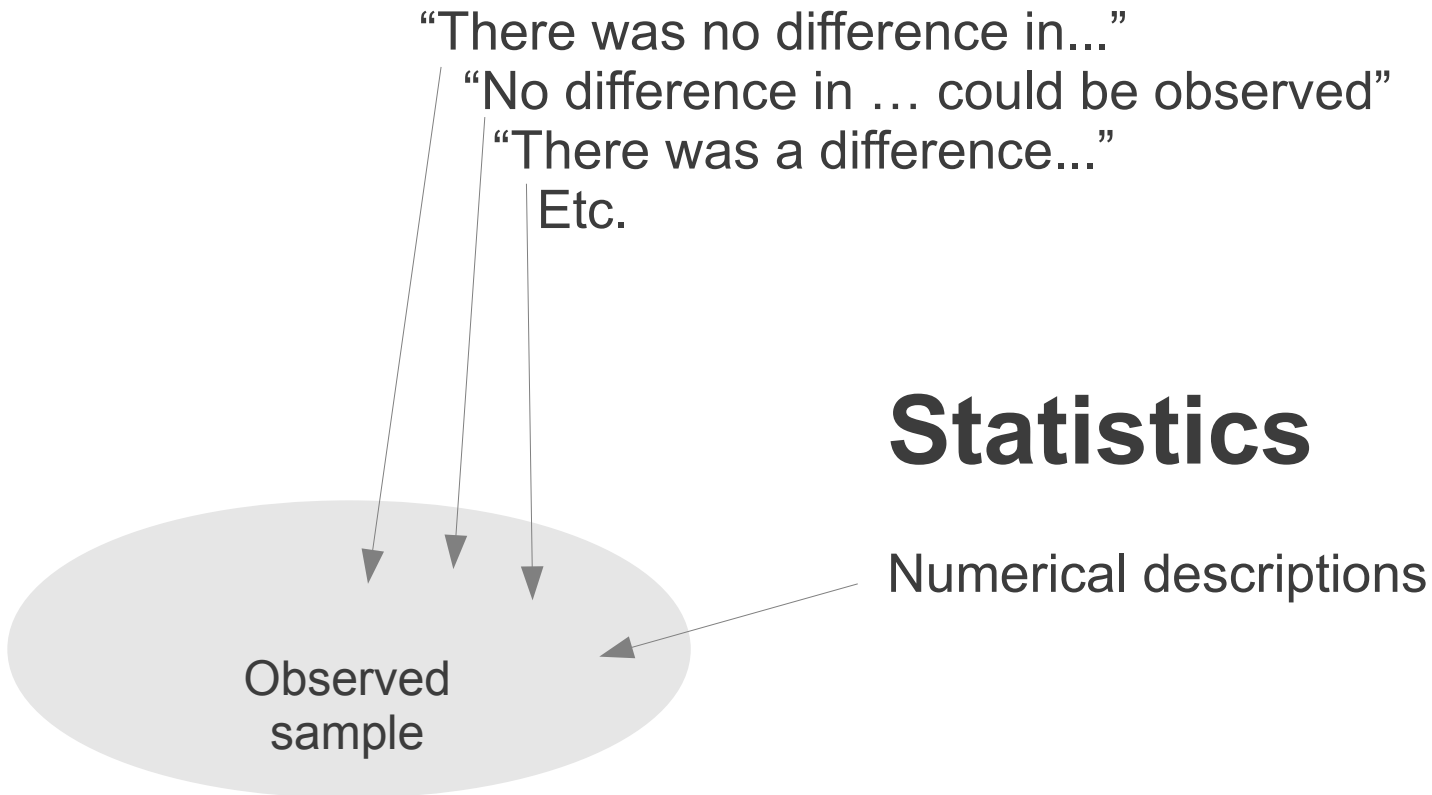
“No difference in ... could be observed”

“There was a difference...”

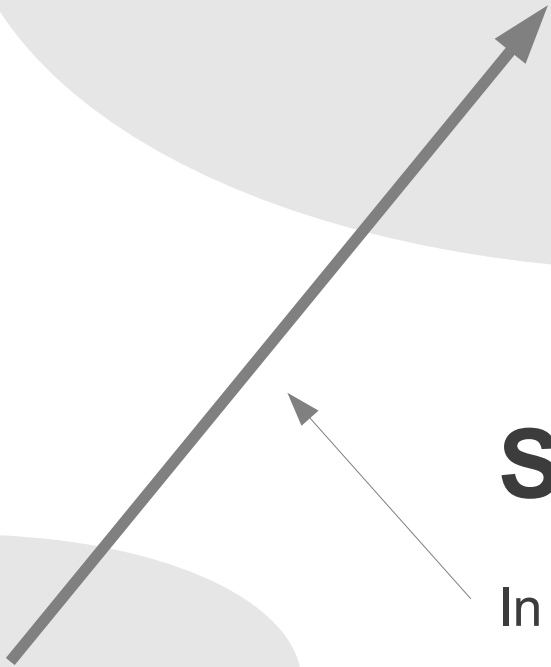
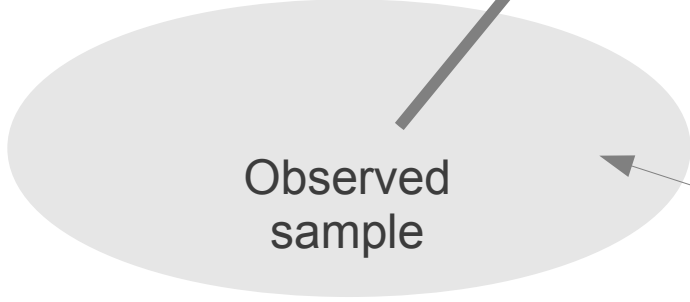
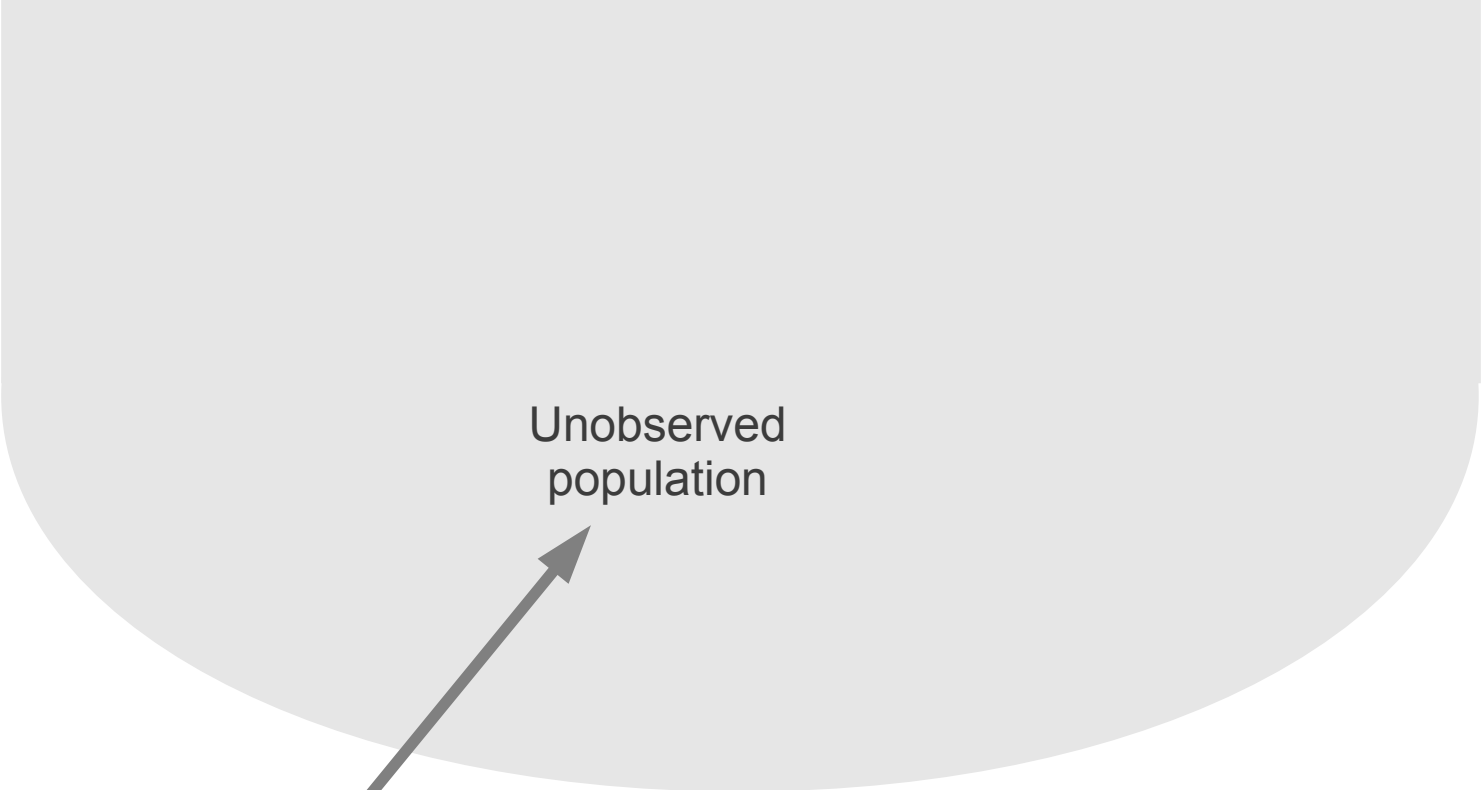
Etc.

## Statistics

Numerical descriptions



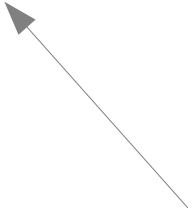
Observed  
sample



# Statistics

In singular: The scientific method of assessing the uncertainty of generalizations

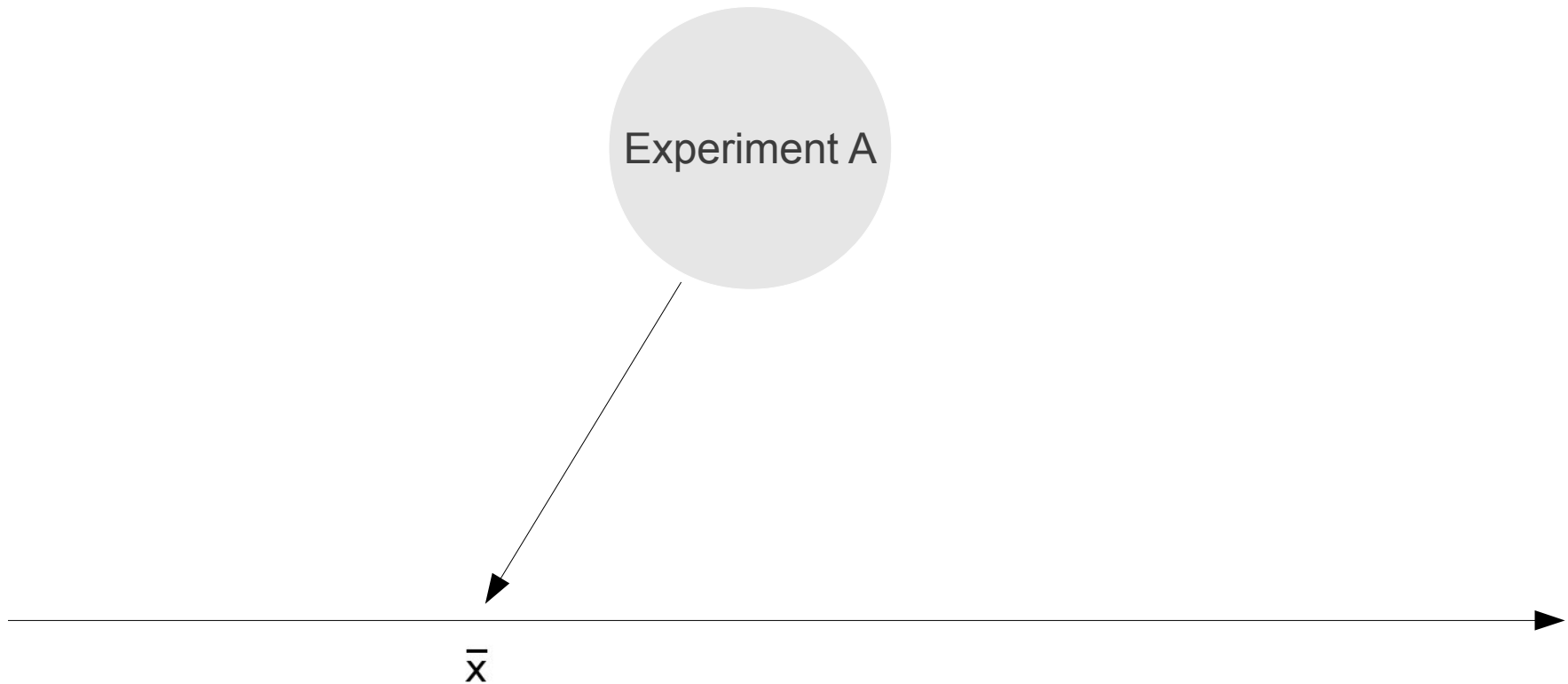
In plural: Numerical descriptions



# Problem

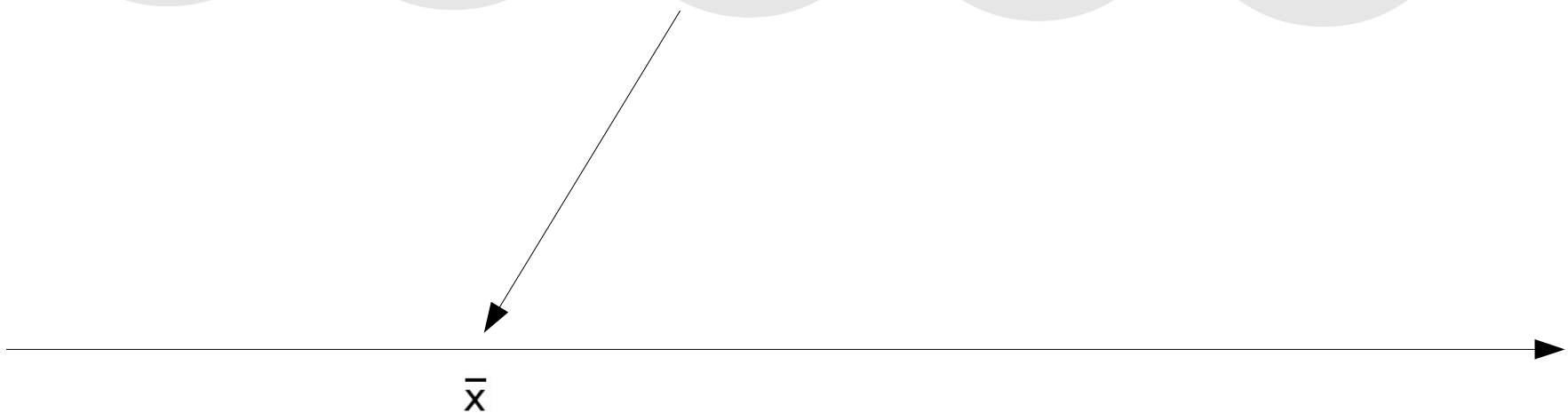
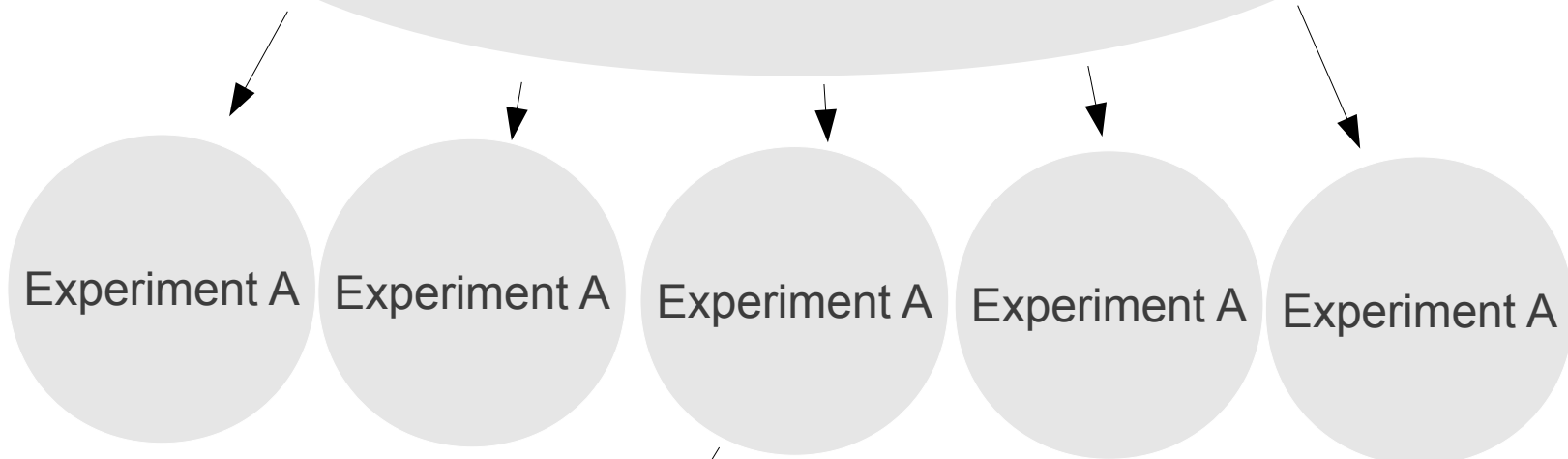
The sample is usually easy to identify, but what “population” are we talking about?

# To what population do experiment A belong?



# To what population do experiment A belong?

The mother of all possible realizations of  
Experiment A



# To what population do experiment A belong?

Experiment A

Experiment A

Experiment A

Experiment A

Experiment A

Experiment A

$\bar{x}$

$\bar{x}$

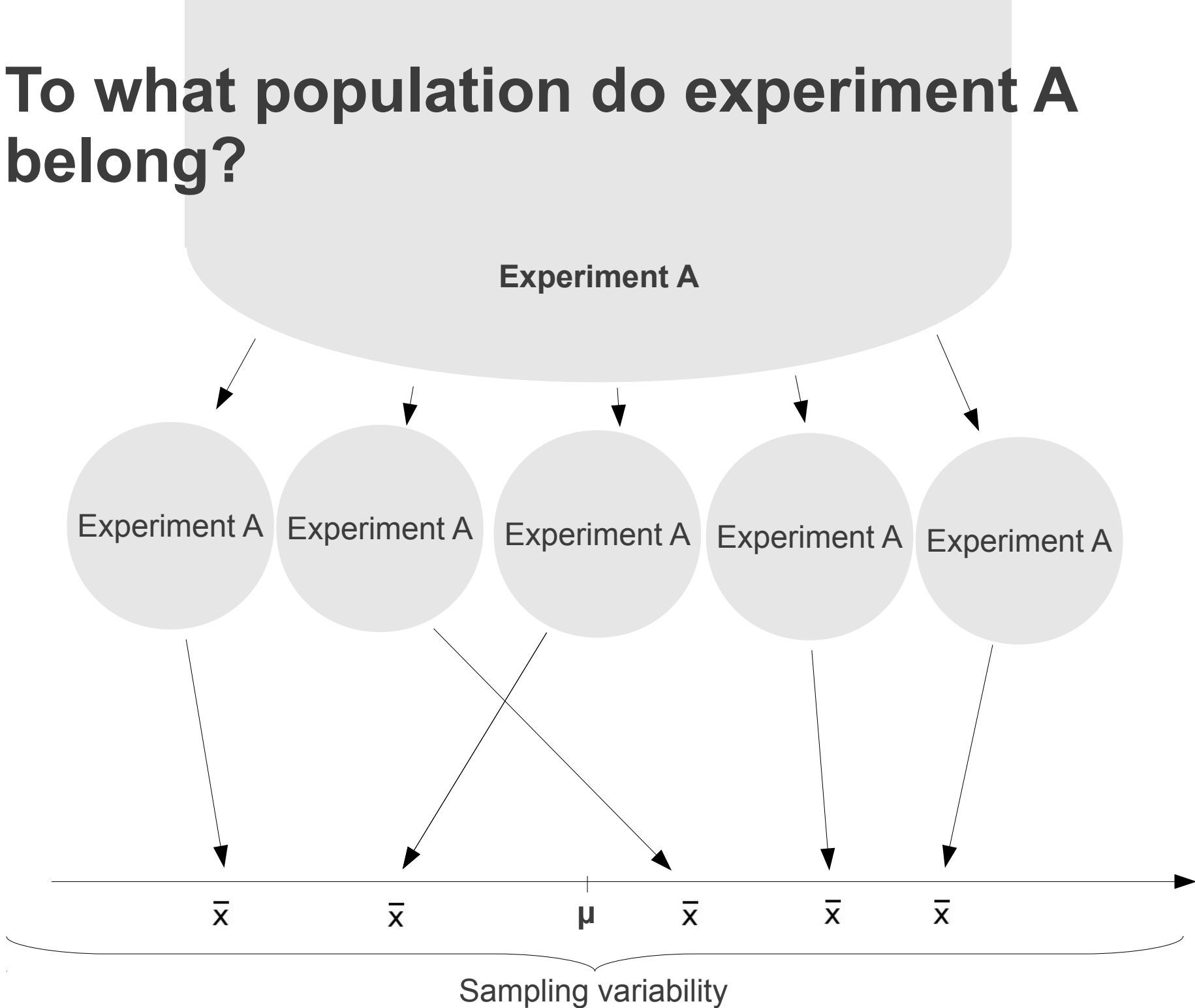
$\mu$

$\bar{x}$

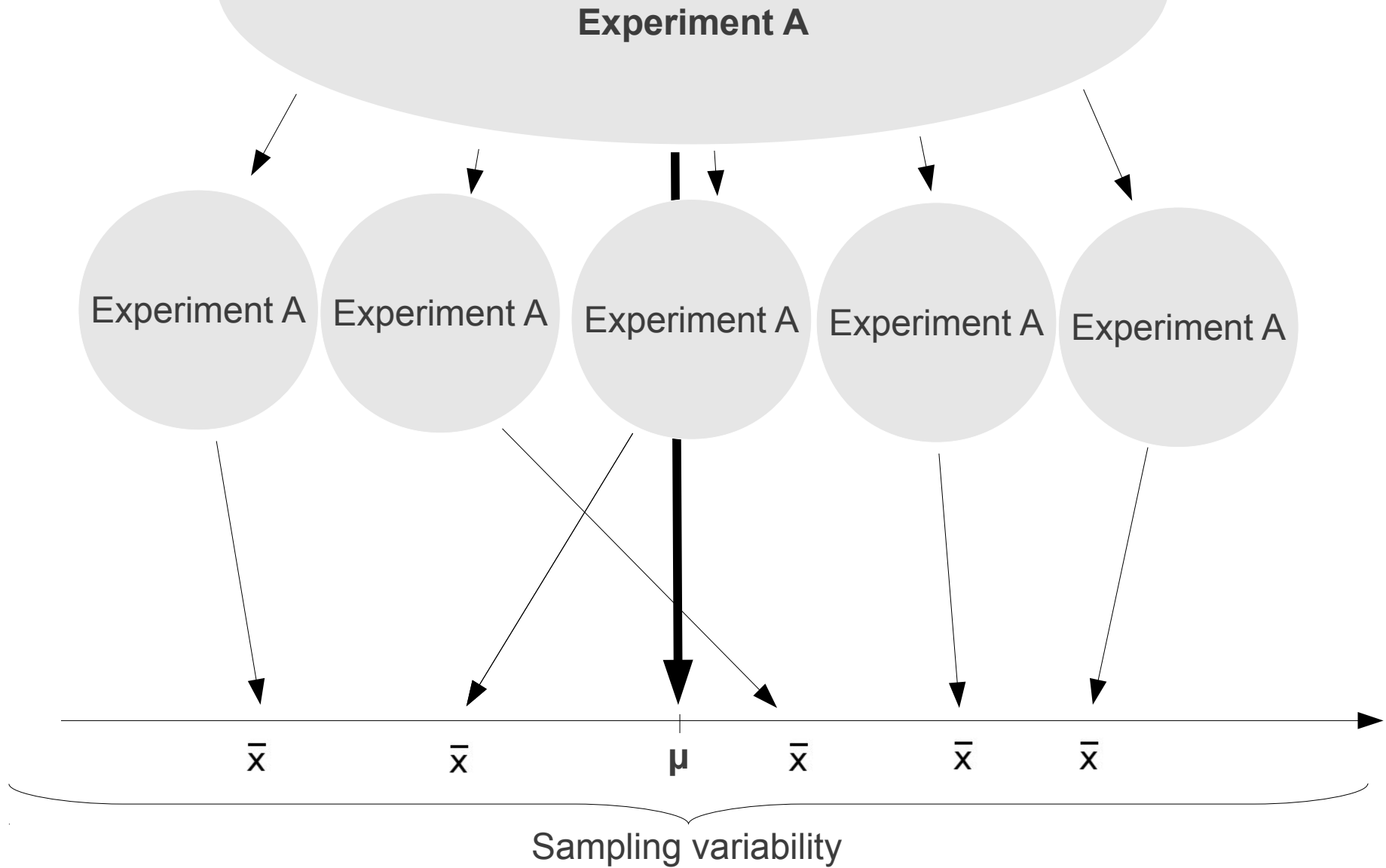
$\bar{x}$

$\bar{x}$

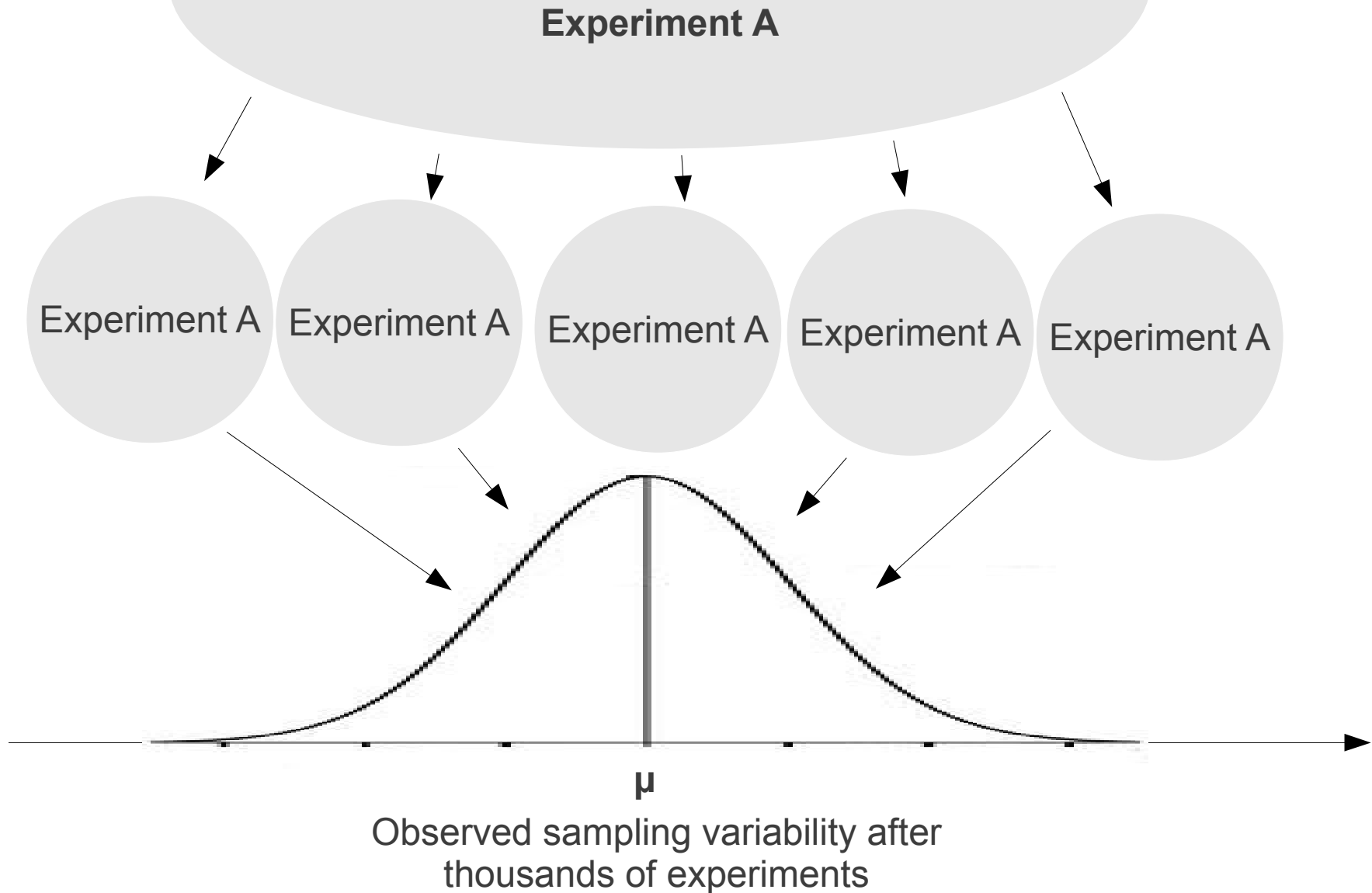
Sampling variability



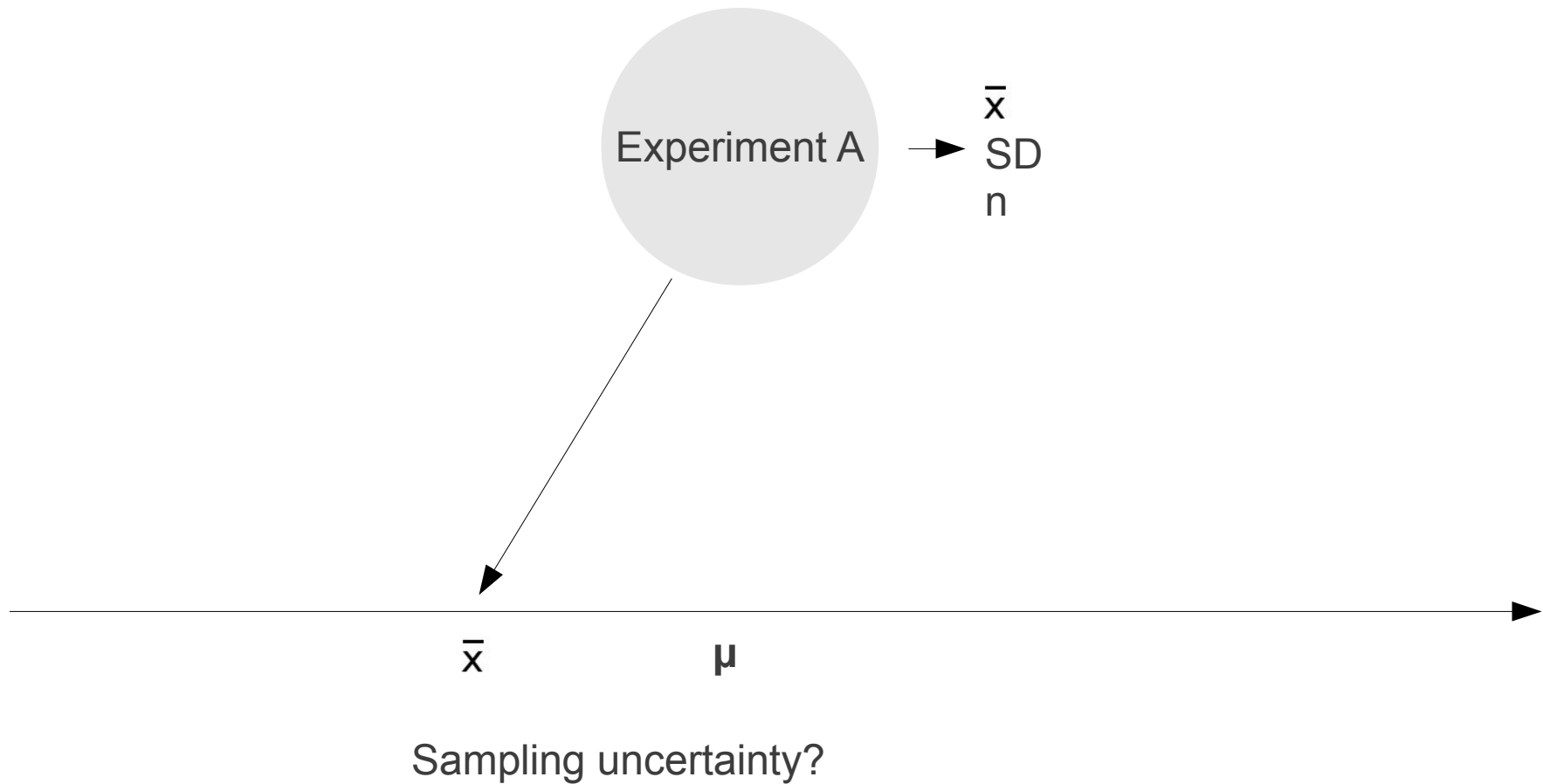
# To what population do experiment A belong?



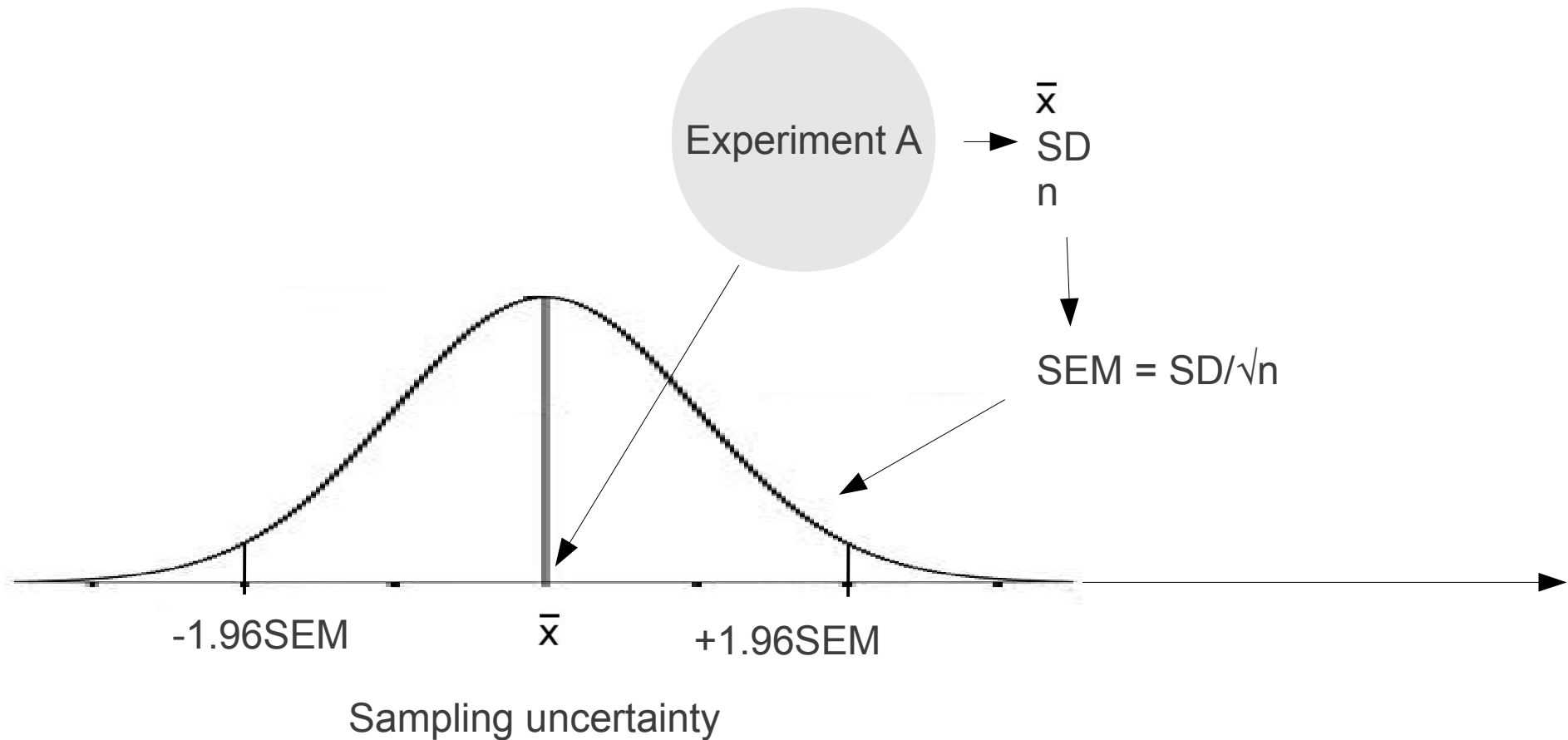
# What is the sampling variability of these experiments?



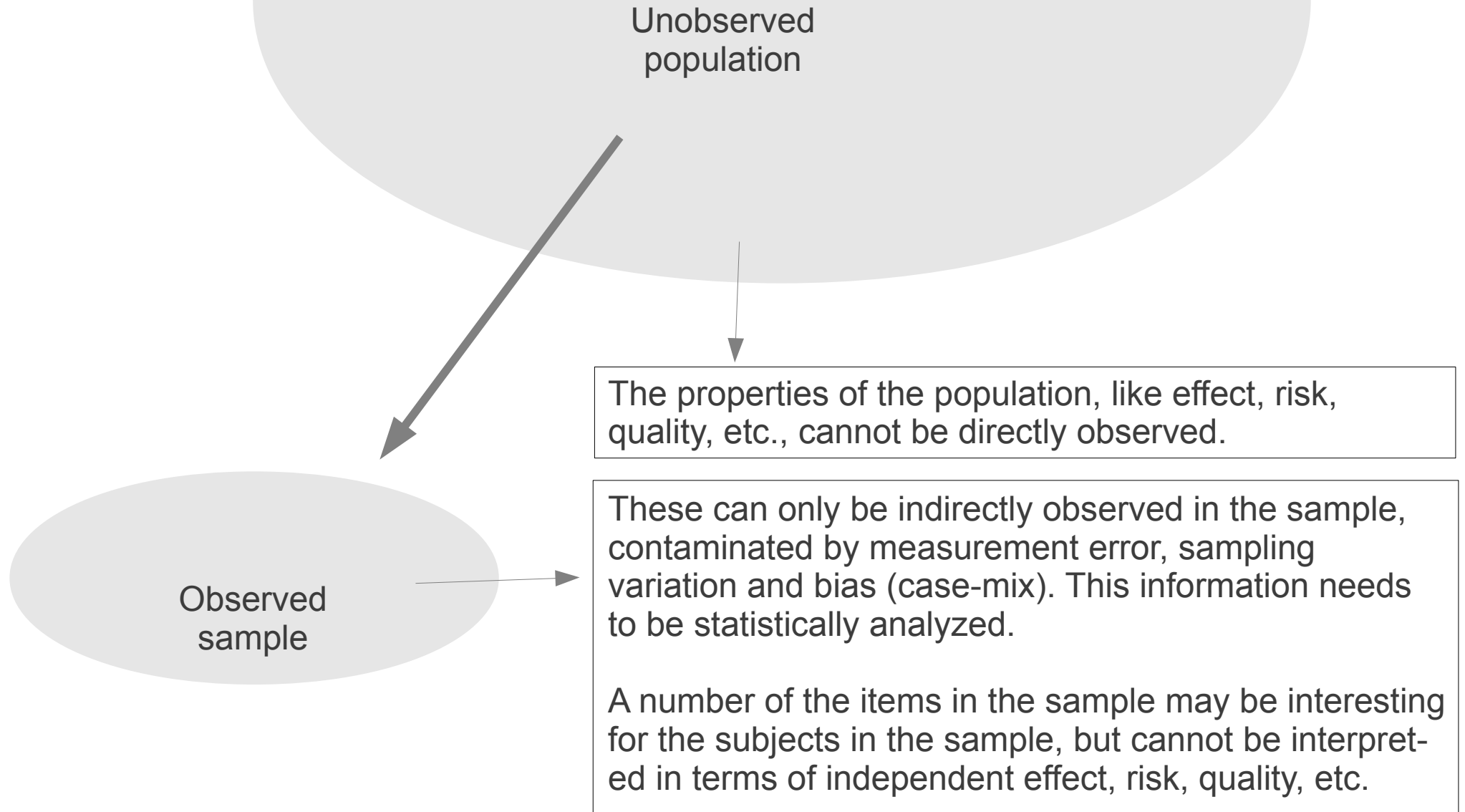
# Can we say anything about sampling uncertainty if only one experiment is performed?



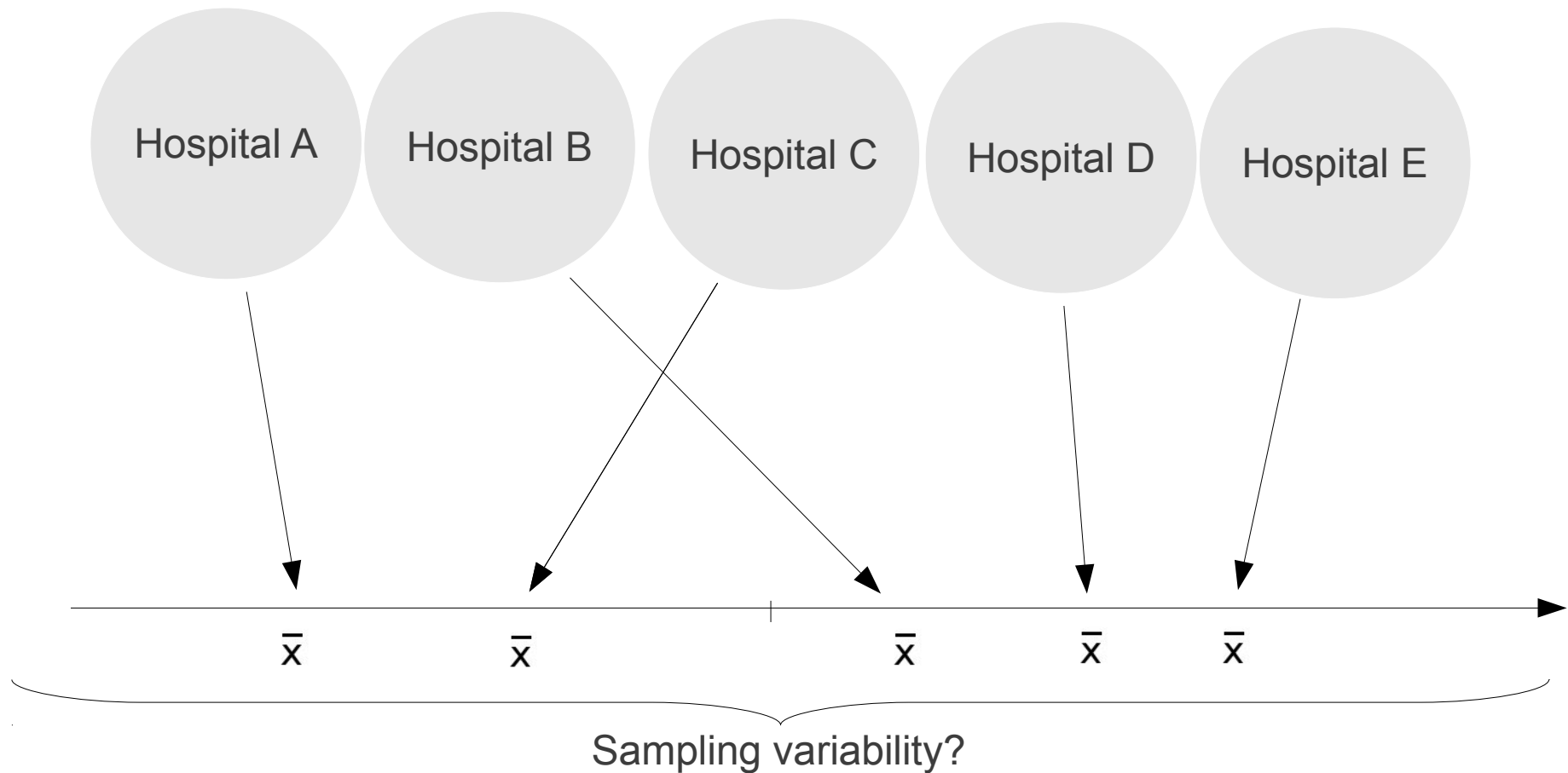
# Can we say anything about sampling uncertainty if only one experiment is performed?



# Is sampling uncertainty a problem only in experimental studies?

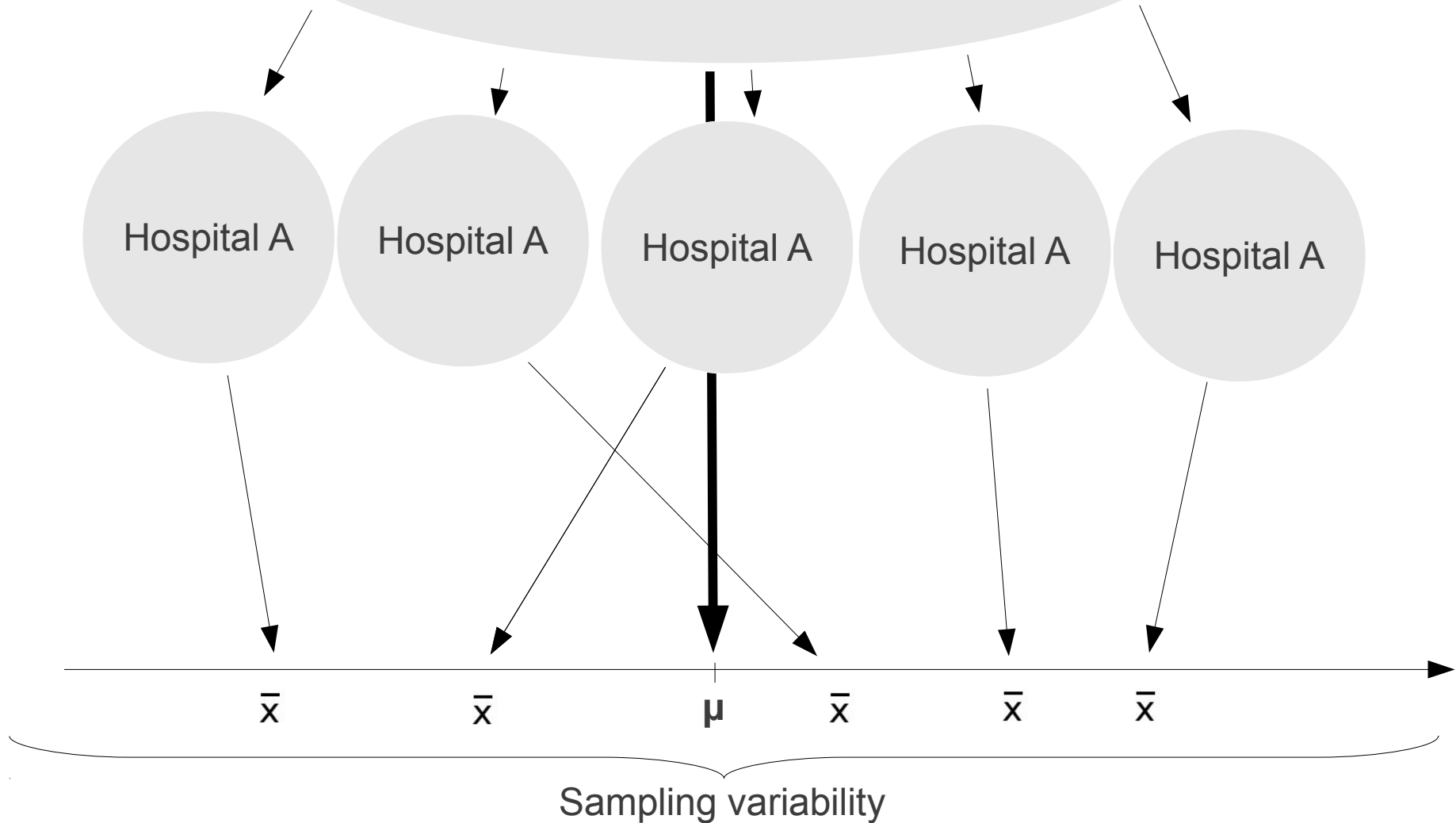


# Do different ranks in league tables represent differences in “hospital quality”?

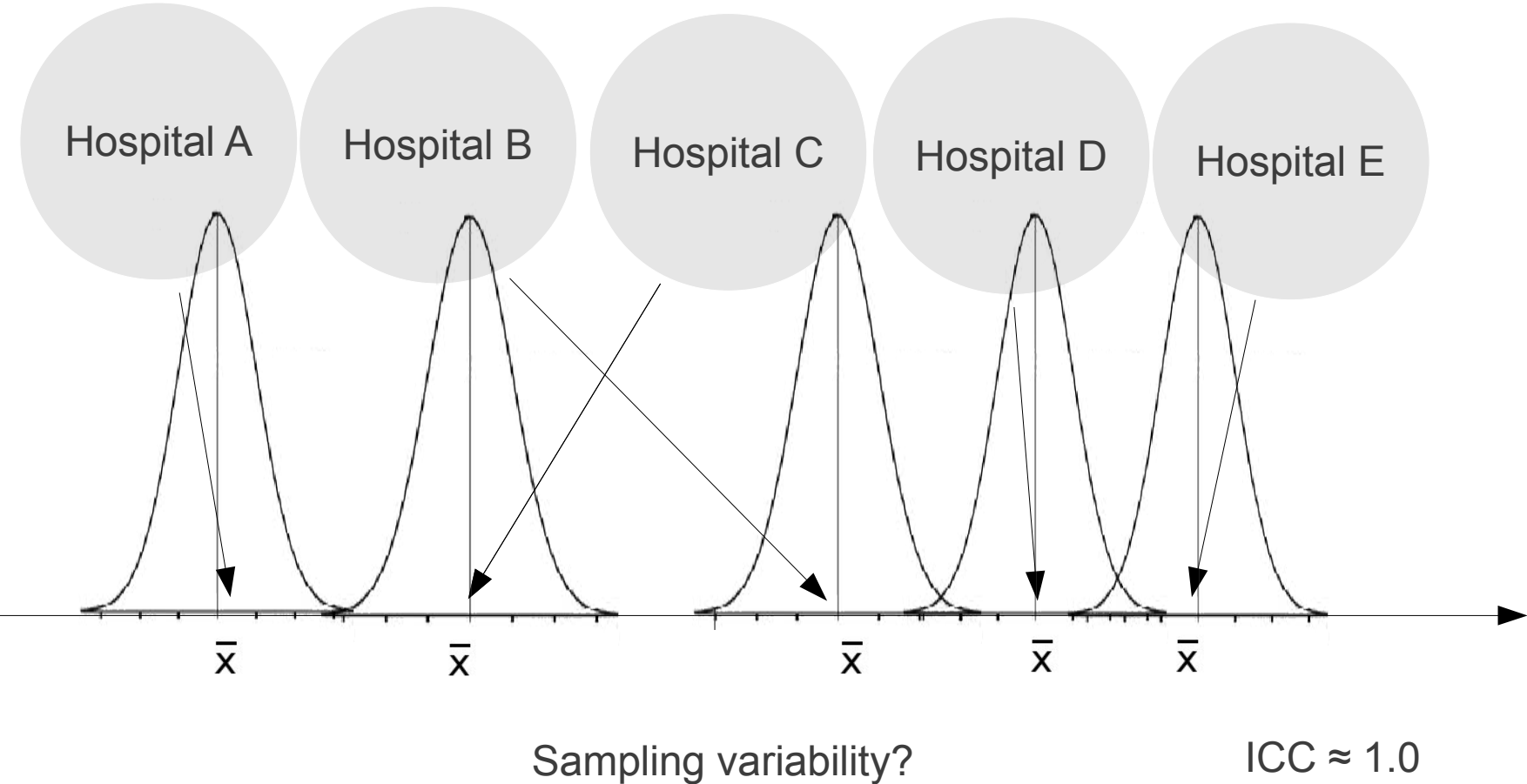


# Or do the differences just reflect sampling variation?

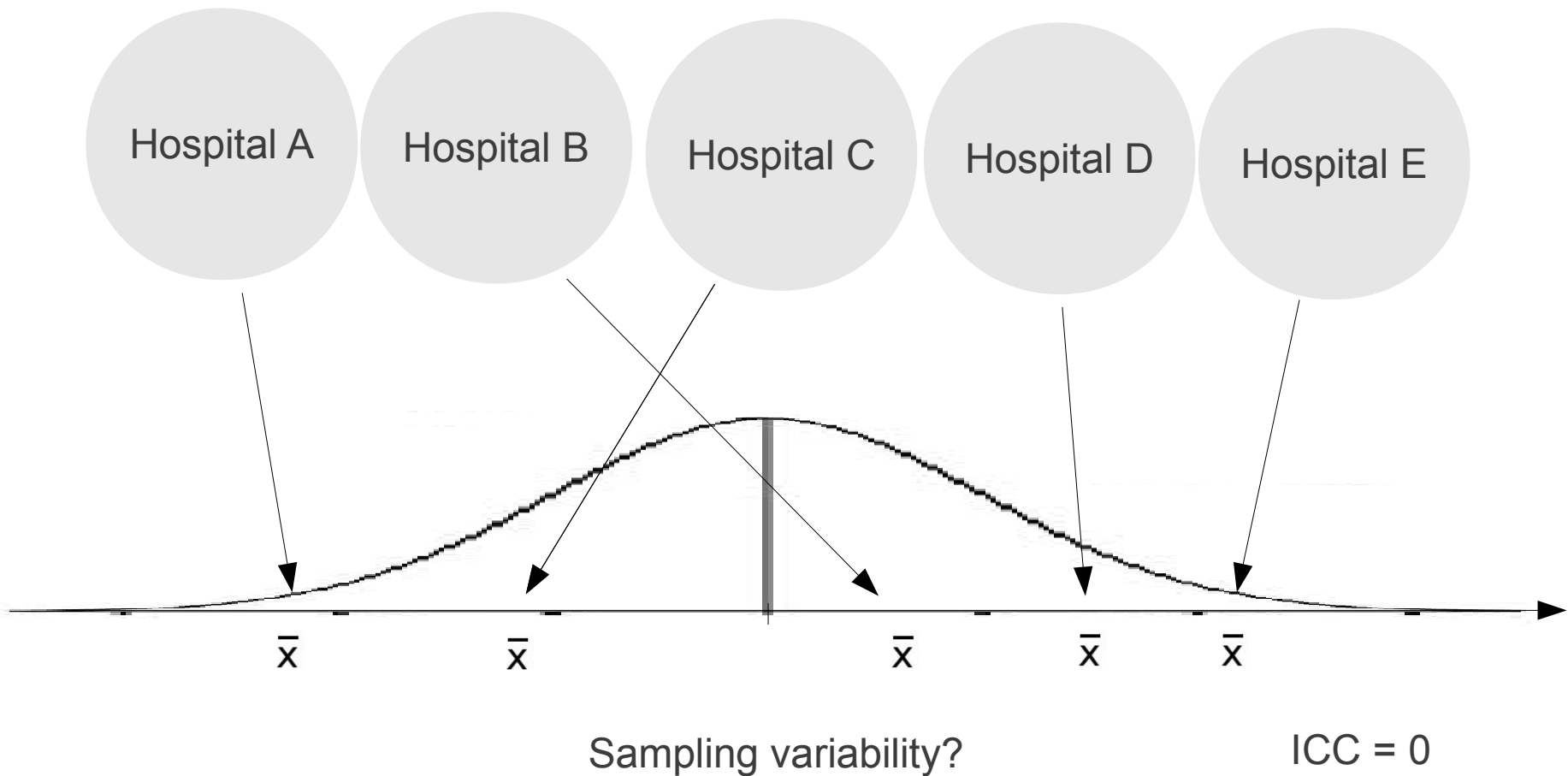
Hospital A



# It depends on the degree of uncertainty!



# It depends on the degree of uncertainty!



# Evaluating uncertainty

## Alt. 1. Hypothesis testing

$$H_0: \mu = 0$$

$$H_A: \mu \neq 0$$

$$P(\bar{\mathbf{X}} | H_0)$$

$P < 0.05 \rightarrow H_0$  is rejected  
statistically significant

# Clinical vs. statistical significance

A clinically significant difference is important for all subjects, whether it is statistically significant or not.

## Example

An increase in systolic blood pressure by 30mmHg.

# Clinical vs. statistical significance

A statistically significant finding is not necessarily clinically significant.

## Example

A decrease ( $p = 0.023$ ) in body weight of 0.1kg.

# Probability

You know, the most amazing thing happened to me tonight. I was coming here, on the way to the lecture, and I came in through the parking lot. And you won't believe what happened. I saw a car with the registration plate ARW 357. Can you imagine? Of all the millions of license plates in the state, what was the chance that I would see this particular one tonight? Amazing...

*Richard P. Feynman*

# Fundamentals of hypothesis testing

The p-value represents the probability of a false positive outcome of a **pre-defined** hypothesis test

Results from testing observed differences (in “fishing expeditions”) are unreliable.

# Fundamentals of hypothesis testing

The probability of a false positive outcome increases with the number of tests performed.

Results from performing multiple tests without recognizing the implicit multiplicity issues are unreliable.

# Fundamentals of hypothesis testing

Presenting **only** the statistical significance of findings is common but should be avoided.

The description “existing” or “not existing” (depending on their p-values)

- misleads the reader about what the investigator actually observed
- says nothing about the clinical relevance of the finding
- is misleading with respect to false positive and false negative findings.

# Fundamentals of hypothesis testing

$$H_0: \mu = 0$$

$$H_A: \mu \neq 0$$

Two-sided test



$$H_0: \mu > 0 \text{ or } \mu = 0$$

$$H_A: \mu < 0$$

One-sided test



Example: The effect of an anti-hypertensive drug.

# Fundamentals of hypothesis testing

$$H_0: \mu = 0$$

$$H_A: \mu \neq 0$$

Two-sided test



$$H_0: \mu > 0 \text{ or } \mu = 0$$

$$H_A: \mu < 0$$

One-sided test



Example: The effect of an anti-hypertensive drug.

Is the null hypothesis clinically meaningful?

# Fundamentals of hypothesis testing

The statistical power to detect a hypothetical difference is used when calculating the required sample size.

The post-hoc power of an observed difference is often calculated but is not meaningful. It does not reveal any other information than the p-value.

The statistical precision of an estimated parameter is best described by its confidence interval.

# Fundamentals of hypothesis testing

Statistical precision depends on the variability of subjects (independent observations) in the population and on the number of observations in the sample.

Testing body parts, e.g. hips, knee, feet, fingers, etc., (intraclass-correlated observations) usually leads to an under-estimation of between-subject variability and to an overestimation of the number of observations.

The consequence is usually too optimistic p-values.

# Fundamentals of hypothesis testing

Tests of imbalance are usually meaningless:

1. baseline imbalance in randomized trials
2. imbalance of matched sets in a matched case-control or cohort study
3. imbalance of exposure related risk factors for use in confounding adjustments.

This imbalance is a property of the sample, and the tests are about properties of the population.

# Evaluating uncertainty

## Alt. 2. Interval estimation

A 95% confidence interval

$$\bar{x} - 2SEM < \mu < \bar{x} + 2SEM$$

Includes with 95% confidence the estimated parameter

# Note!

$\bar{X} \pm 3SEM$       99.7% confidence interval

$\bar{X} \pm 2SEM$       95% confidence interval

$\bar{X} \pm 1SEM$       68% confidence interval

# Note!

$\bar{X} \pm 3SEM$  99.7% confidence interval

$\bar{X} \pm 2SEM$  95% confidence interval

$\bar{X} \pm 1SEM$  68% confidence interval

$\bar{X} \pm 1SD$  is a measure of observed dispersion

# Note!

$\bar{X} \pm 3SEM$  99.7% confidence interval

$\bar{X} \pm 2SEM$  95% confidence interval

$\bar{X} \pm 1SEM$  68% confidence interval

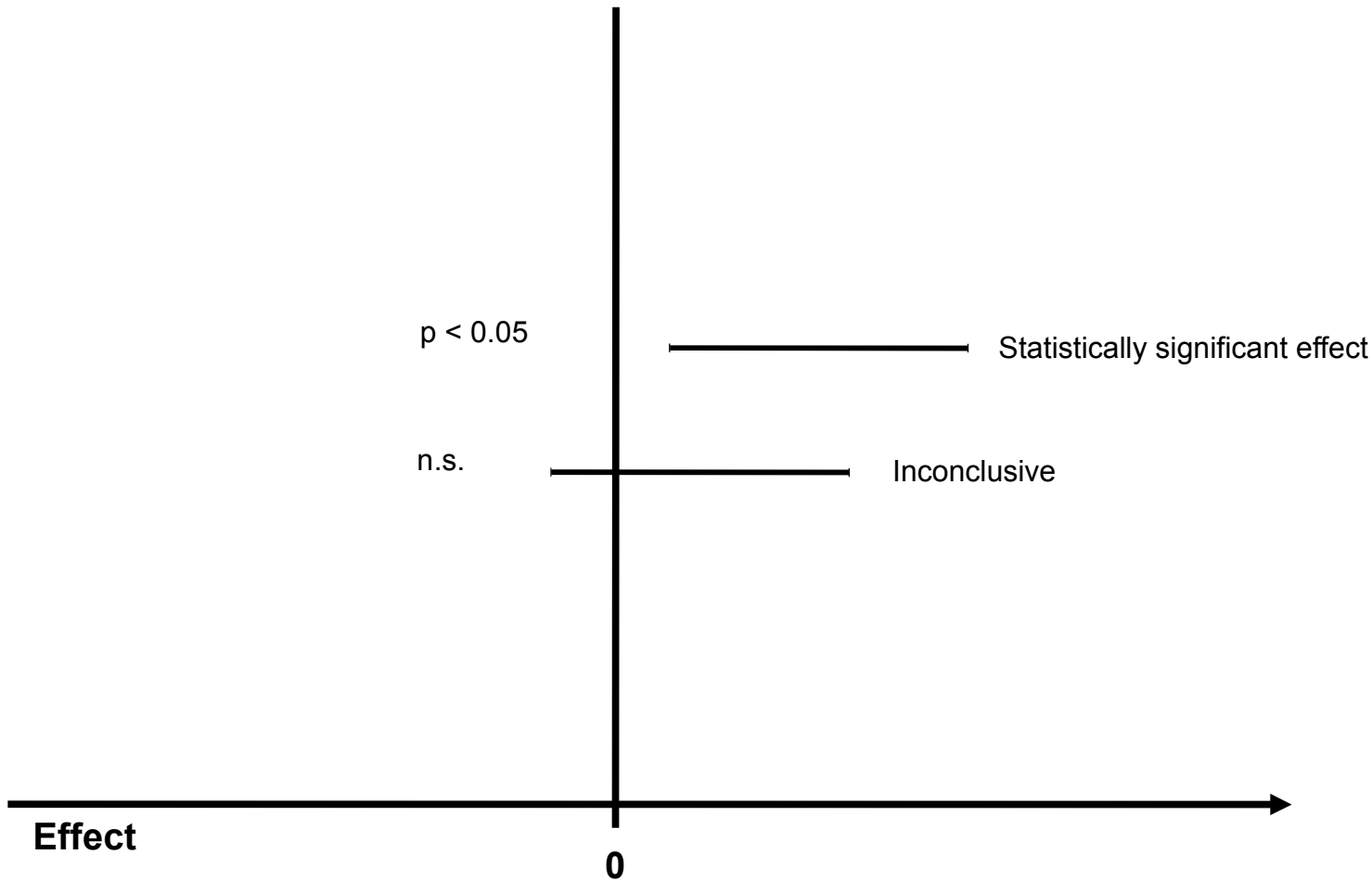
$\bar{X} \pm 1SD$  is a measure of observed dispersion

$\pm SD \approx 95\%Ci$  for the mean when  $n = 6$

# P-value and confidence interval

Information in p-values  
[2 possibilities]

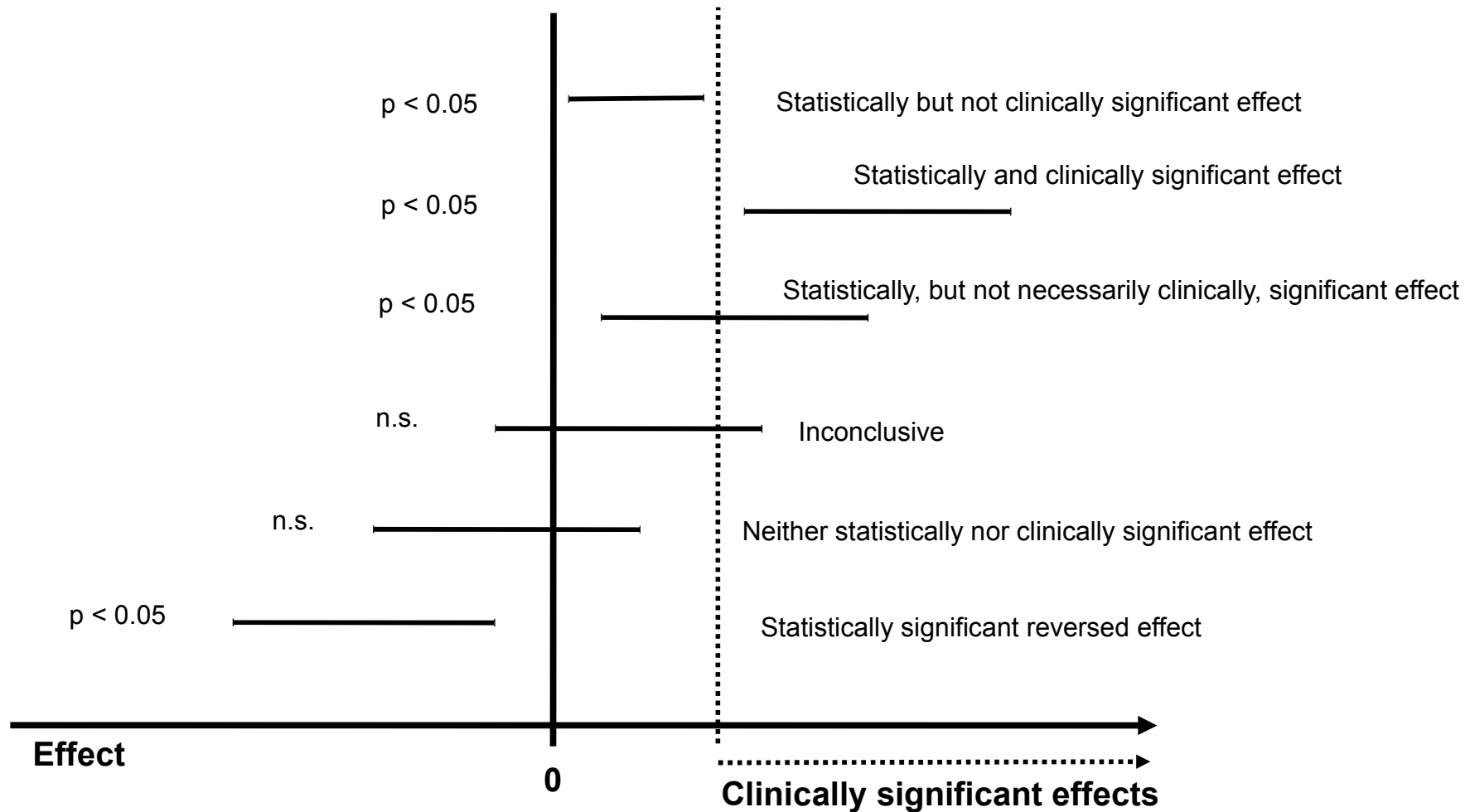
Information in confidence intervals  
[2 possibilities]



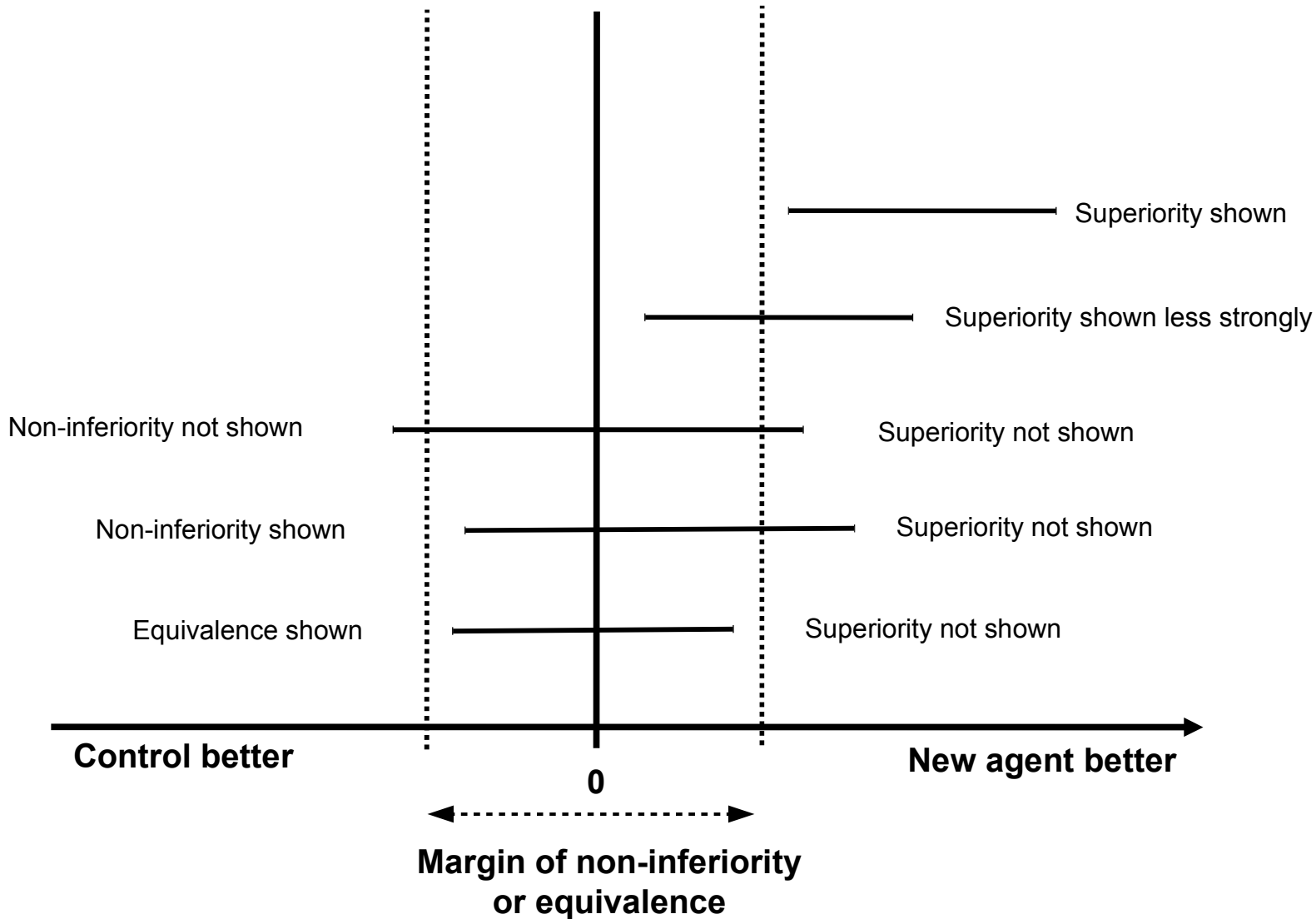
# P-value and confidence interval

Information in p-values  
[2 possibilities]

Information in confidence intervals  
[6 possibilities]



# Superiority vs. non-inferiority



# Experimental vs. observational studies

**Experiments:** Bias is eliminated by design:

“Block what you can, randomize what you cannot”.

Statistical analysis: Protect the type-1 error rate

**Observation:** Blocking and randomization is impossible.

The results must be adjusted in the statistical analysis.

Statistical analysis: Prioritize validity

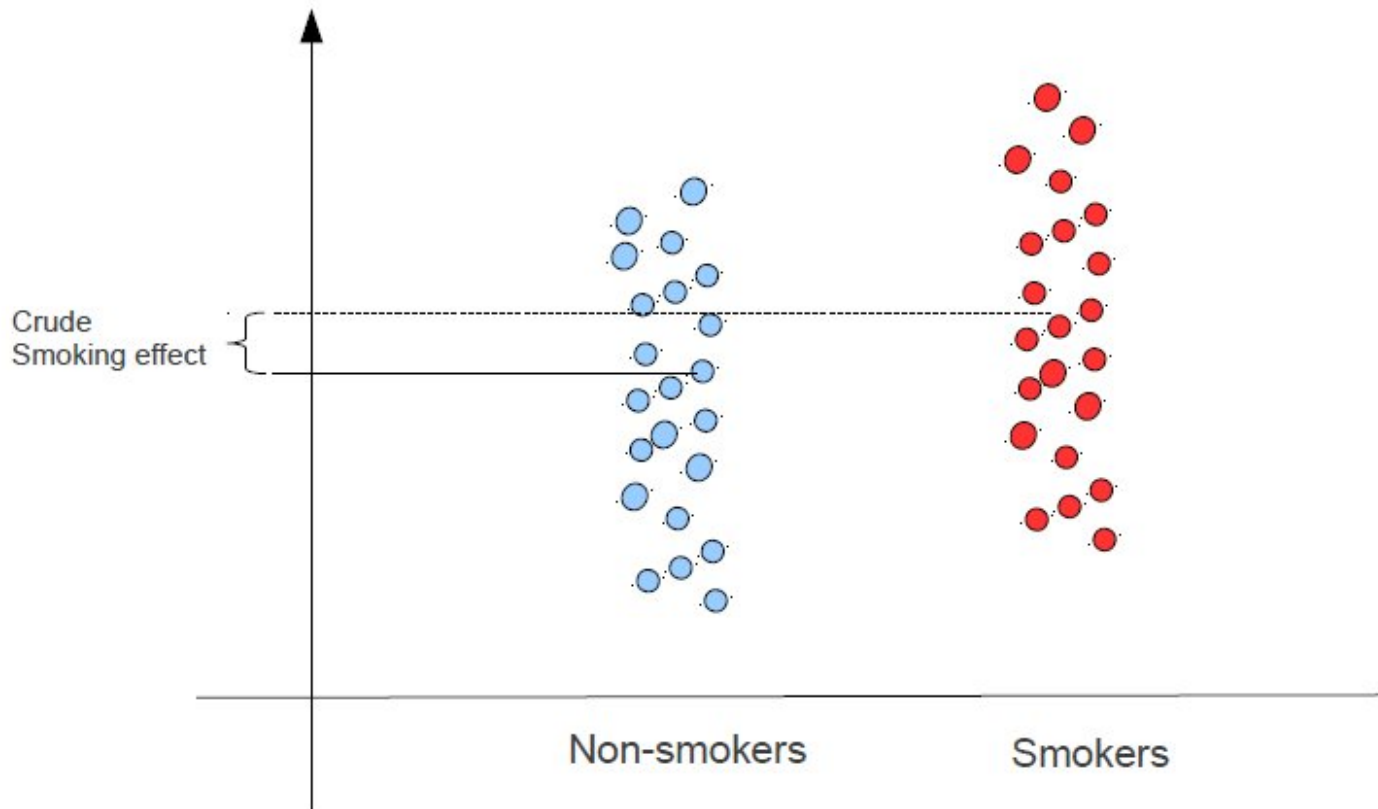
# Observational studies

## Validity

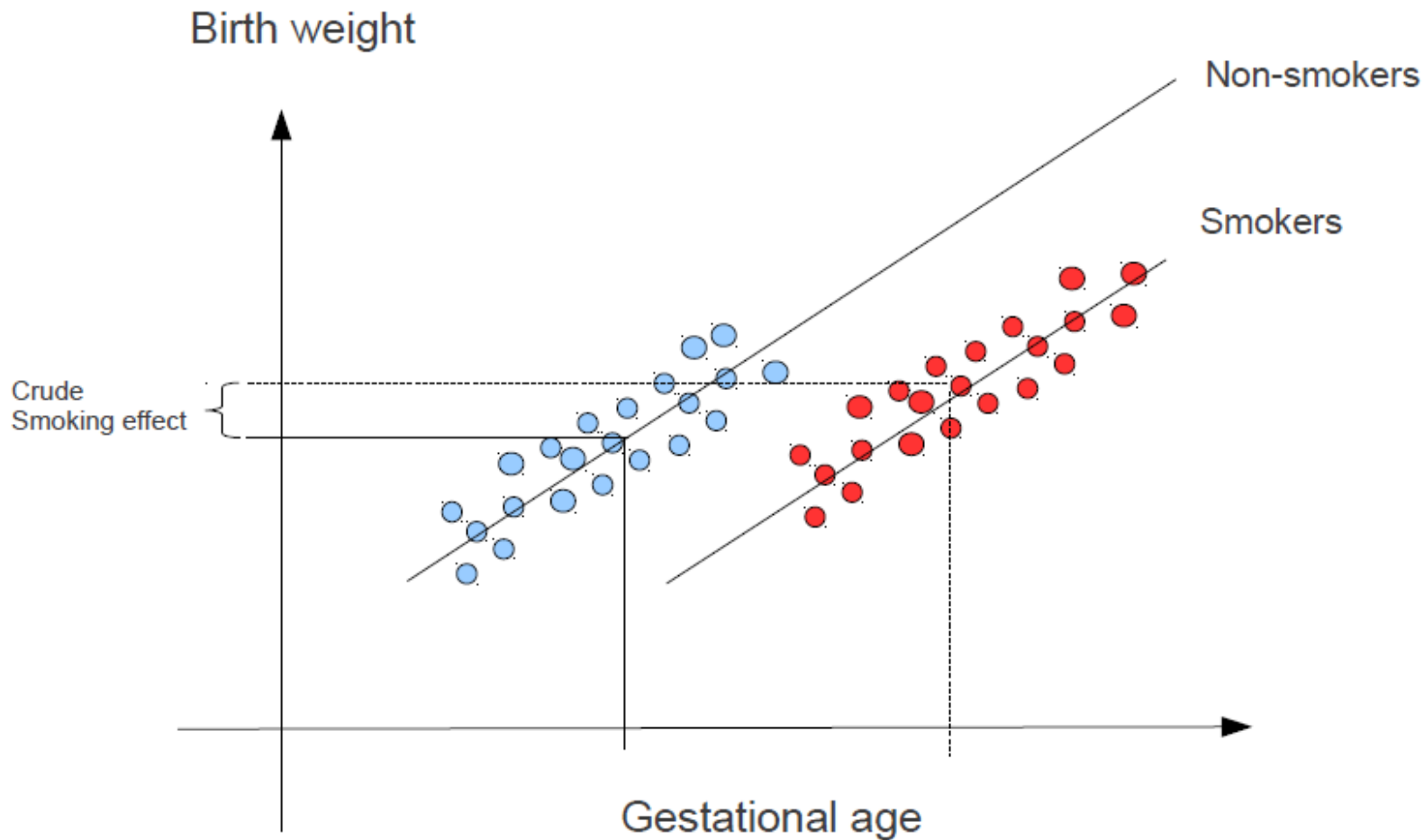
- Selection bias (systematic differences between comparison groups caused by non-random allocation of subjects)
- Information bias (misclassification, measurement errors, etc.)
- Confounding bias (inadequate analysis, flawed interpretation of results)

# Confounding bias – crude estimate

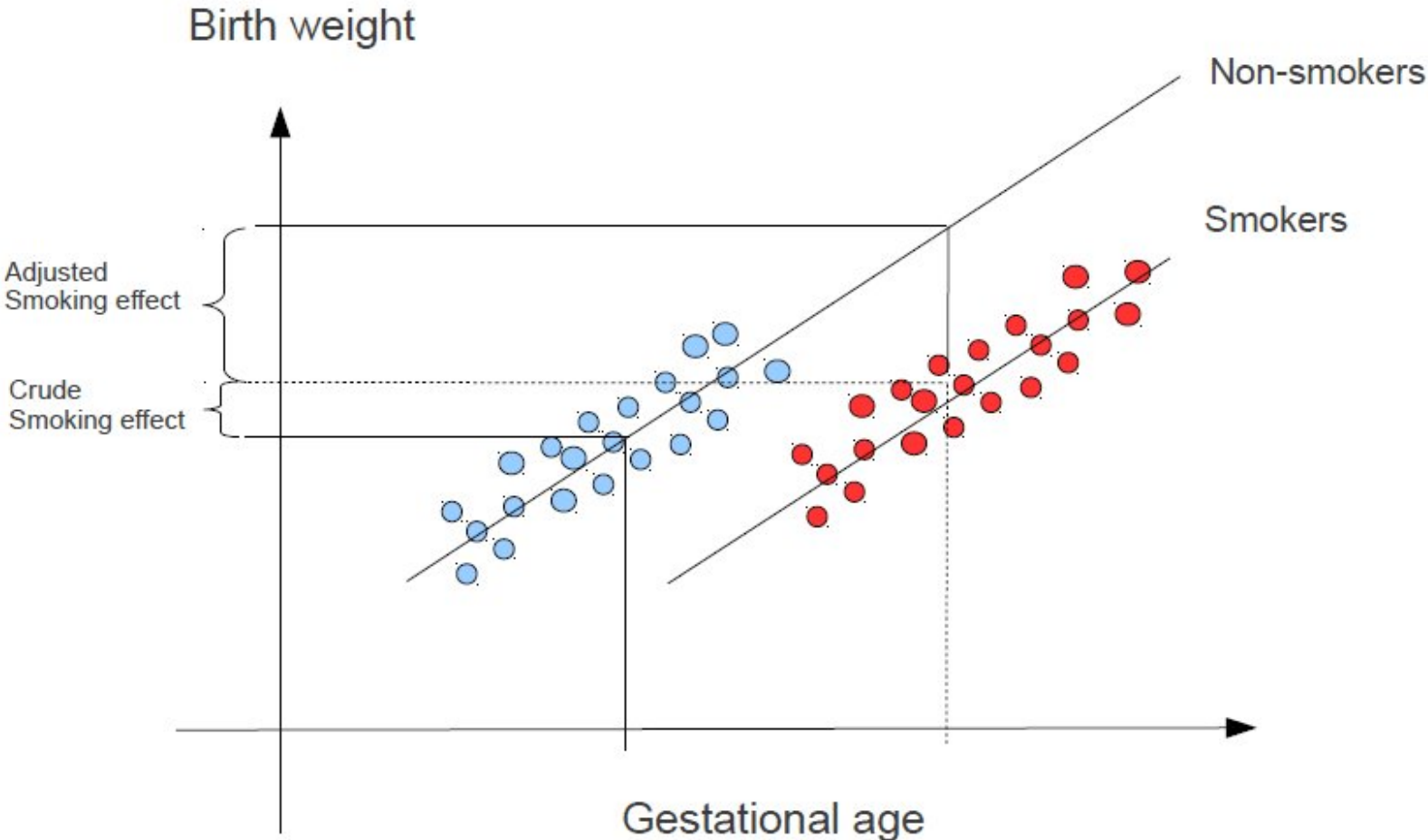
Birth weight



# Confounding bias – crude estimate



# Confounding bias – adjusted estimate



# Testing for confounding

Univariate screening for statistically significant effects, or stepwise regression, is often used to select covariates for inclusion in a regression model.

Confounding bias is a property of the sample, not of the population. What relevance have hypothesis tests?



"I'M SORRY BUT THERE ARE NOW 16,000 MEDICAL JOURNALS, AND I NO LONGER HAVE TIME TO SEE ANY PATIENTS."

**Thank you for your attention!**