Sampling and uncertainty

Michael Otterstatter

BCCDC Biostats Session

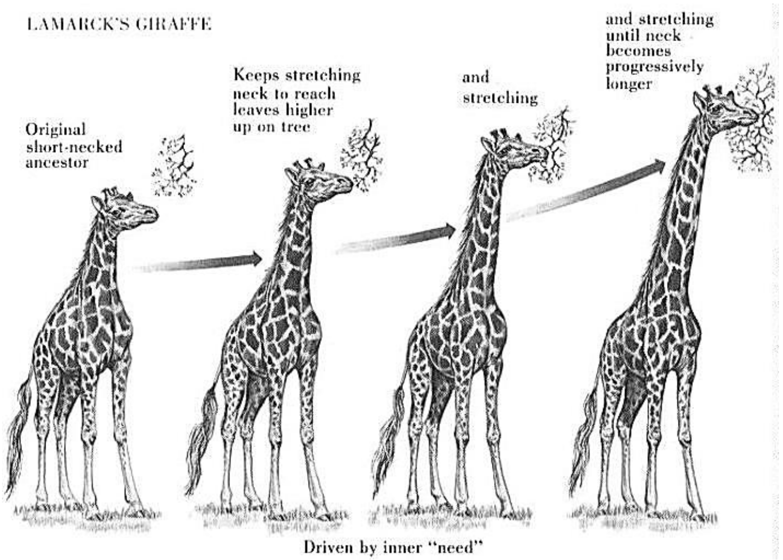
June 7, 2019

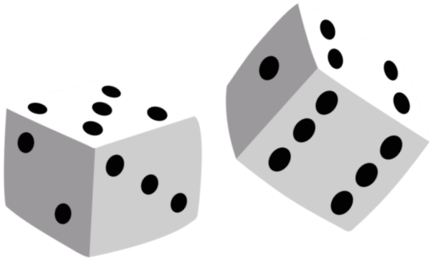
Session Overview

- In this session we will discuss
    - concepts of variability and uncertainty
    - how these concepts apply to population sampling
    - how these concepts apply to statistical models

Background

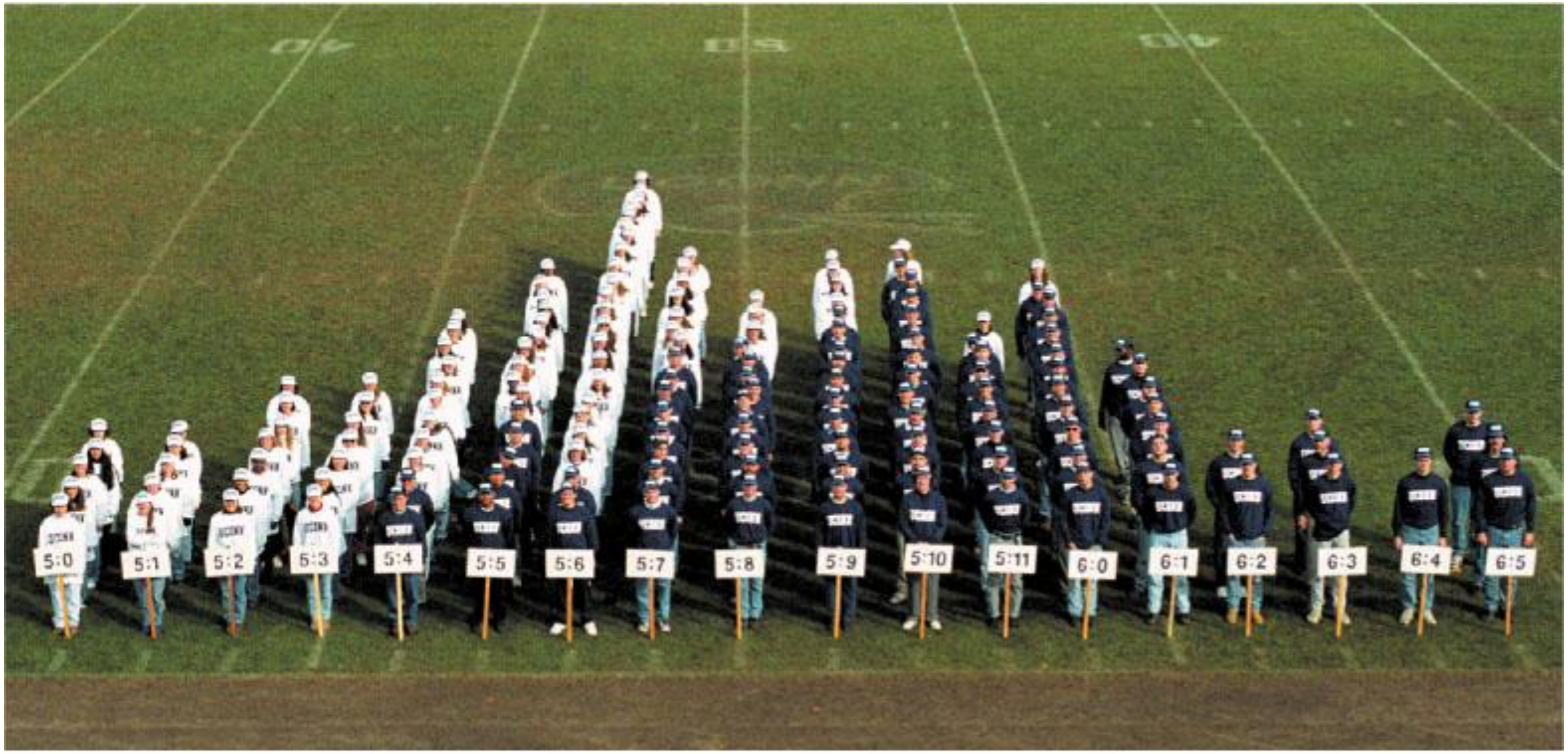
- Two key concepts
  - **Natural variability** – differences between individuals or groups (arising from genetic and/or environmental differences)
  - **Uncertainty** – lack of precise knowledge of characteristics, processes or events (arising from randomness in nature, or incomplete information)





Background

- **Natural variability** is a feature of the natural world, a quantity of interest that we wish to measure



Background

- **Uncertainty** is a nuisance that we wish to remove – however, observations (data) almost always have uncertainty, either because of

  - incomplete information or disagreement regarding a *knowable* true value (e.g., measurement error, missing data, etc.)

  or,

  - inherent unpredictability of an *unknowable* true value (e.g., randomness, complexity, etc.)

Background

- Uncertainty can be reduced by collecting more and better information, but never entirely removed
  - there is always measurement error, even if very small
  - uncertainty due to randomness cannot be reduced

Populations and samples

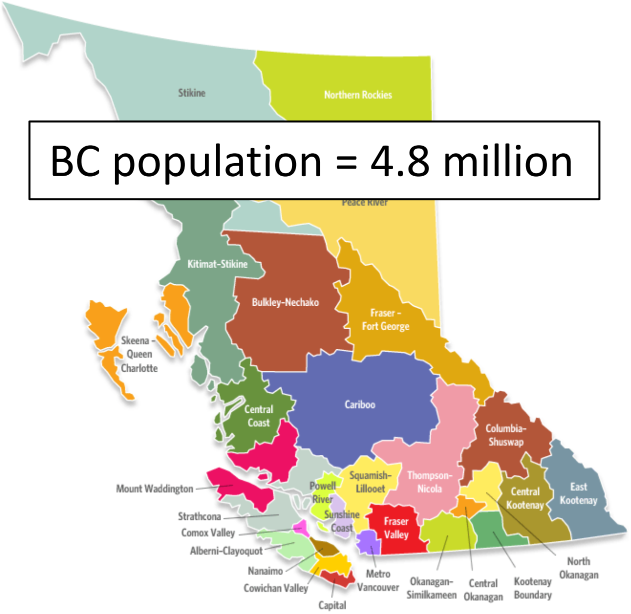
- Typically our goal is to understand health-related phenomena in a large group of individuals (population)

- Two options are available:
  1. observe/measure every individual in the population (**rare**)
  2. use statistics to infer from samples to population (**often**)

Populations and samples

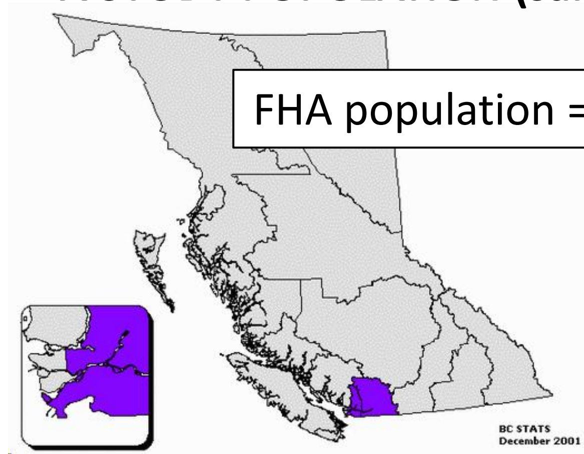
A TARGET POPULATION

BC population = 4.8 million



A STUDY POPULATION (sampling frame)

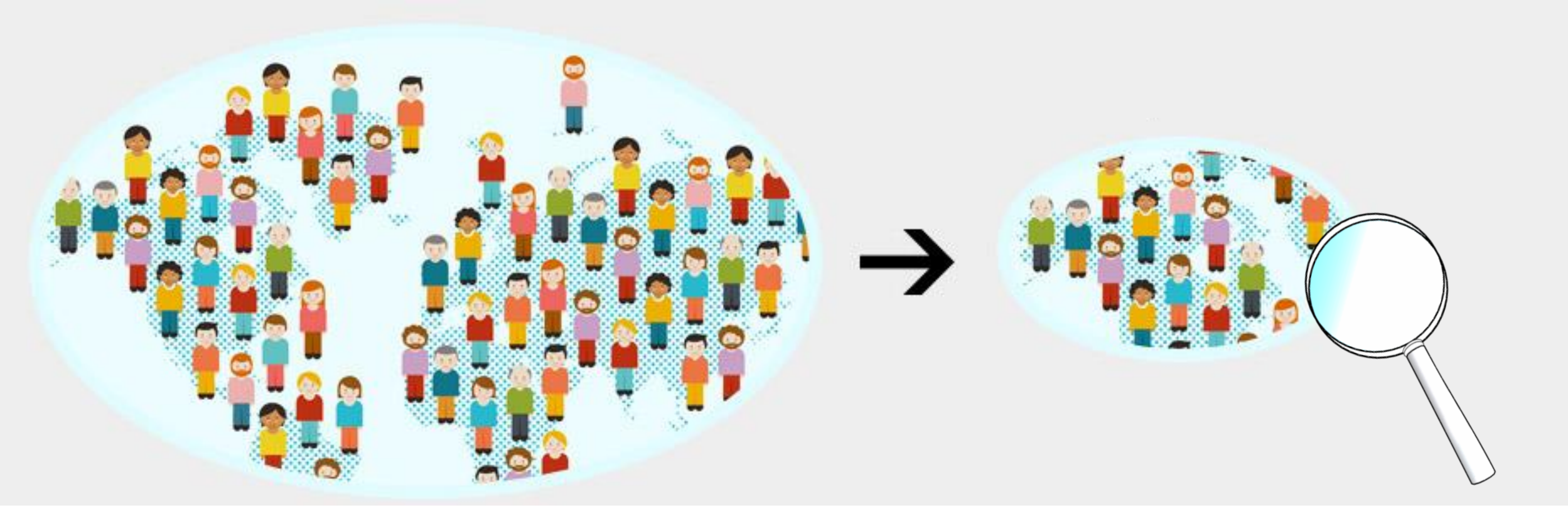
FHA population = 1.7 million



A STUDY SAMPLE

Populations and samples

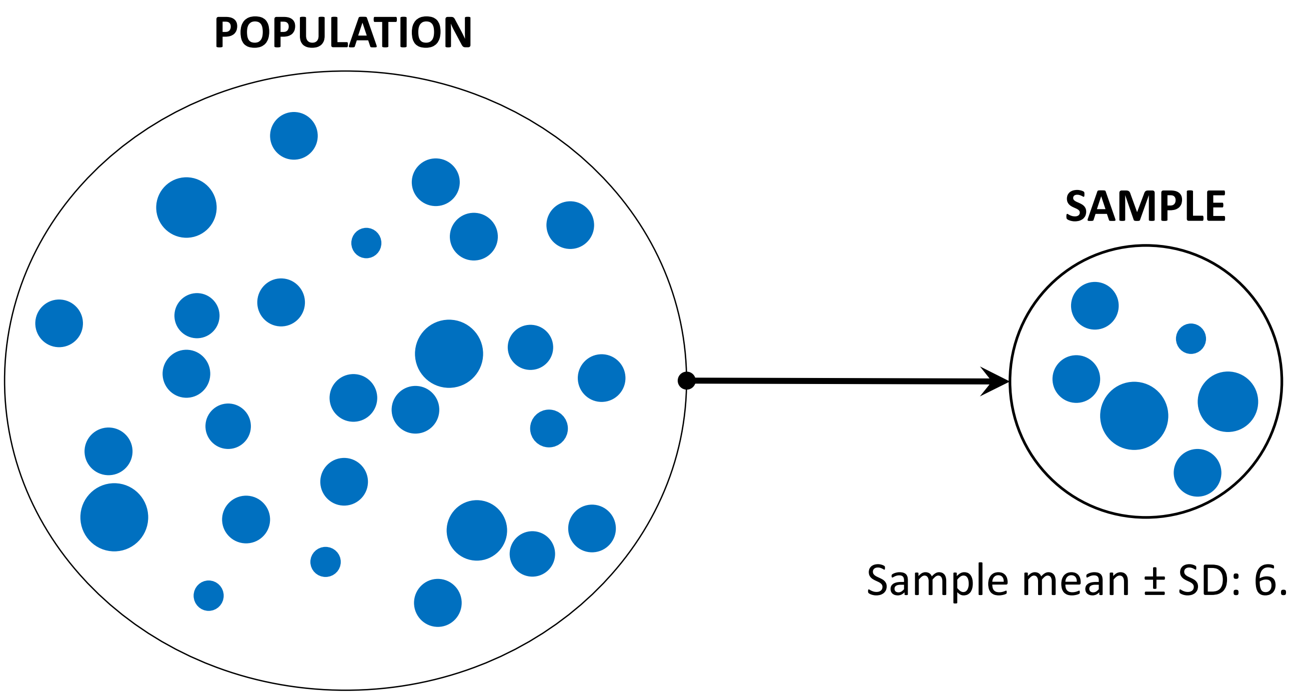
Population Sample



- Taking measurements from population samples has important implications for variability and uncertainty

Populations and samples

- Natural variability in a population can be estimated though representative samples and *measures of variation*



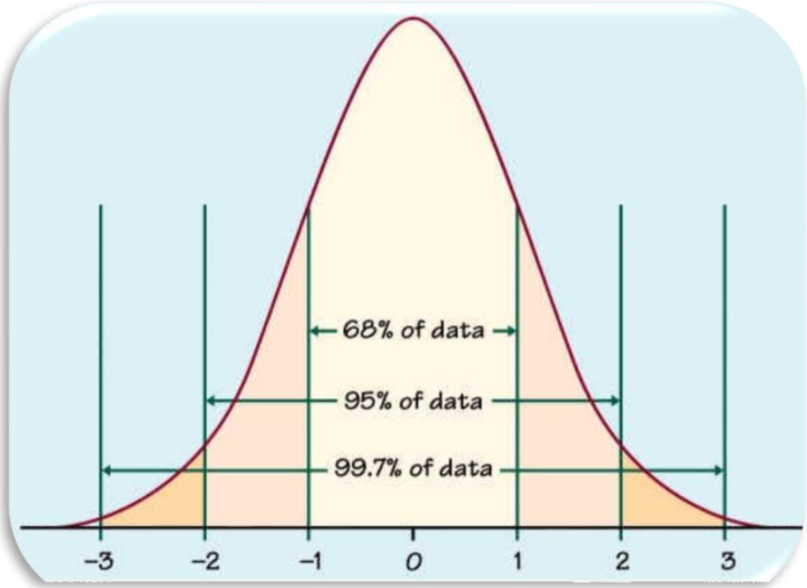
Sample mean ± SD: 6.9 ± 1.8

Population mean ± SD: 6.2 ± 1.6

Measures of variation

- **Variance**: how spread out data are in a sample (or in a population)
  - average squared deviation of data points from the mean
- **Standard deviation**: spread of data around the mean in a sample (or a population)
  - square root of variance

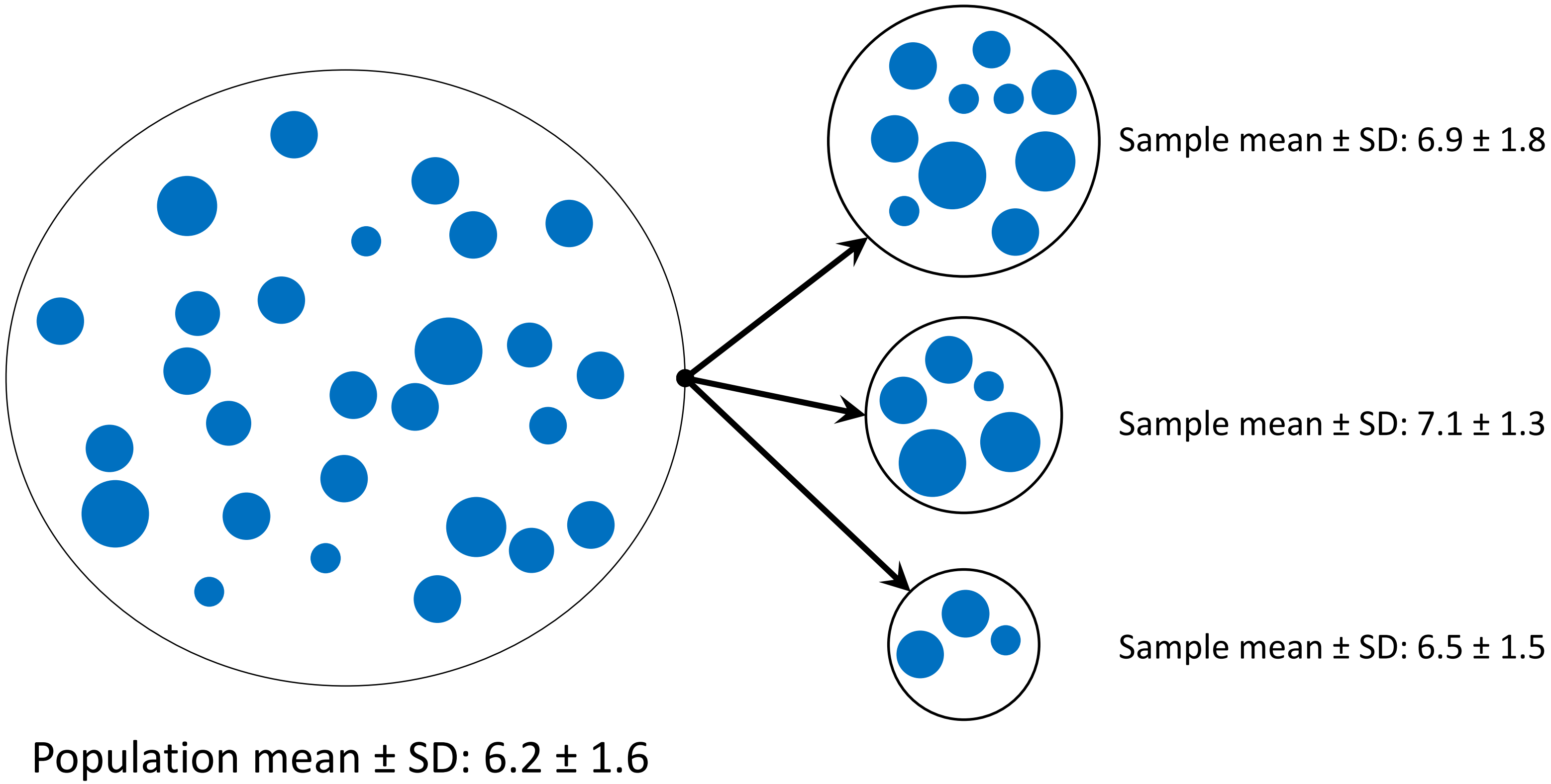
For normally distributed data, mean and SD determine data intervals



Standard deviations from the mean

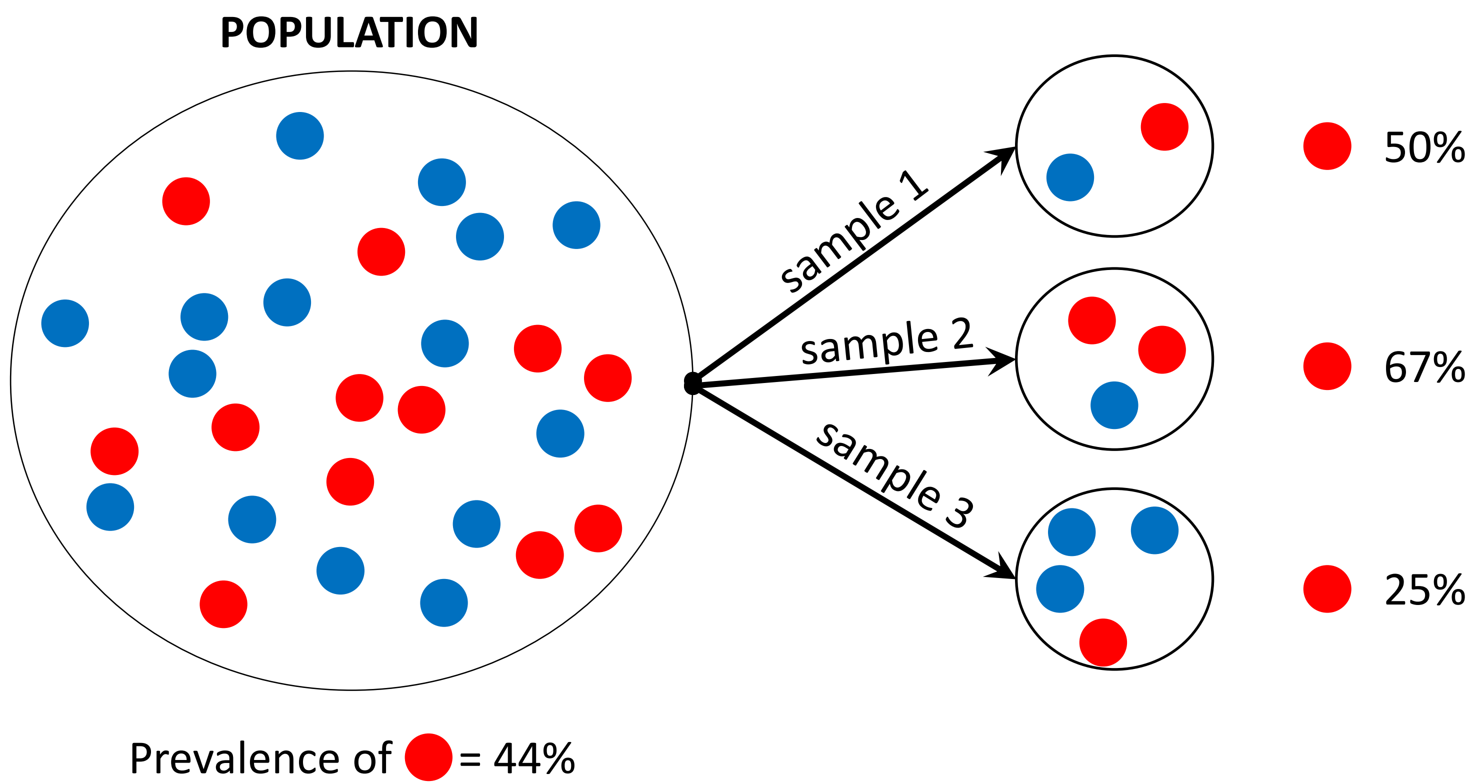
Measures of variation

- Can we reduce natural variability in our data?
- If we increase our sample size, does that change our measures of variation?



Populations and samples

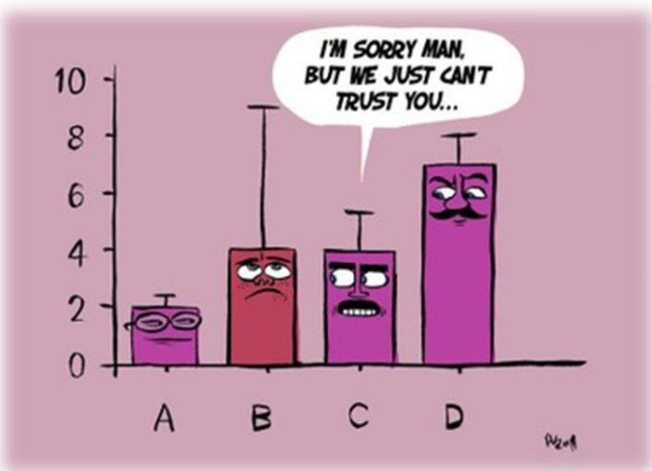
- However, samples do not necessarily behave alike, or exactly like the population -- hence, we have **sampling error** and *measures of uncertainty*



BC Centre for Disease Control
An agency of the Provincial Health Services Authority

Measures of uncertainty

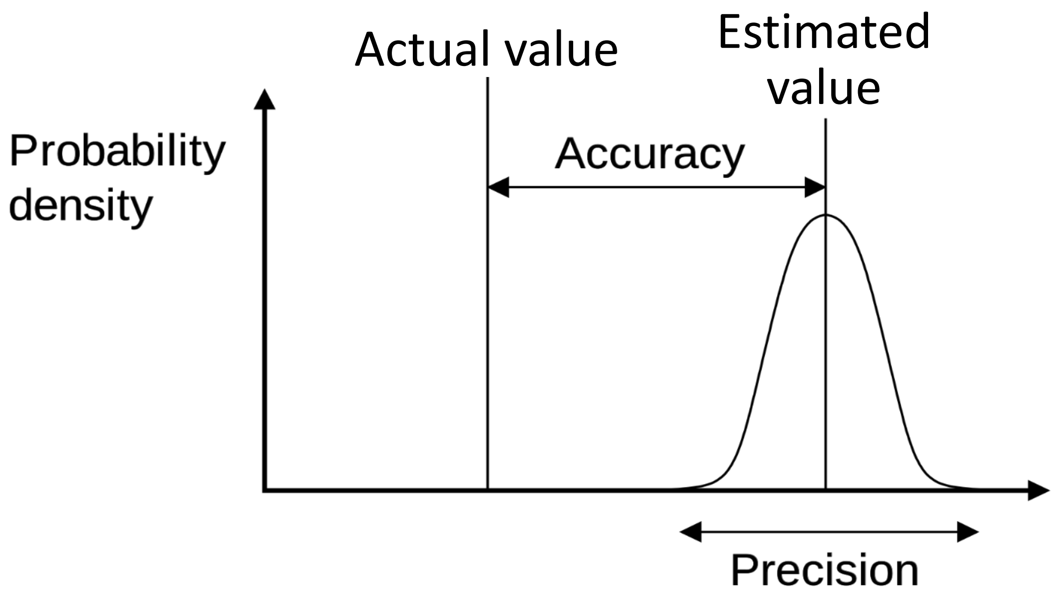
- Repeatedly sampling a population generates a distribution of values (means, for example)
  - **Standard error**: the SD of this distribution; a measure of uncertainty or *precision*
  - **Confidence intervals**: interval of this distribution within which a sample statistic will fall (e.g., sample mean falls within this interval 95% of the time)

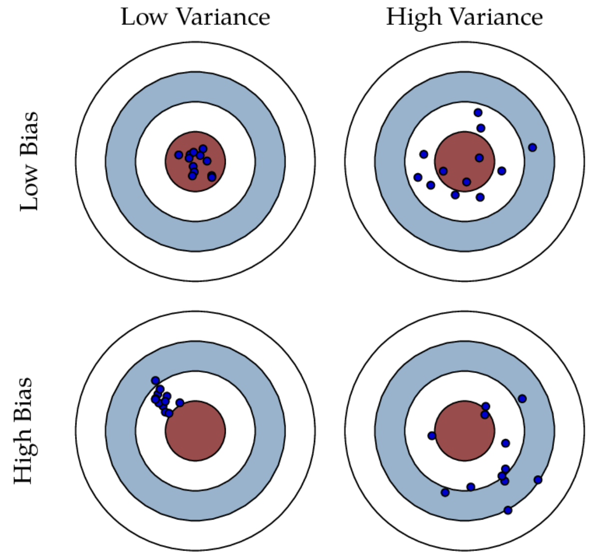


BC Centre for Disease Control
An agency of the Provincial Health Services Authority

Measures of uncertainty

- Note that measures of uncertainty reflect the *precision* of study results, not necessarily the *accuracy*





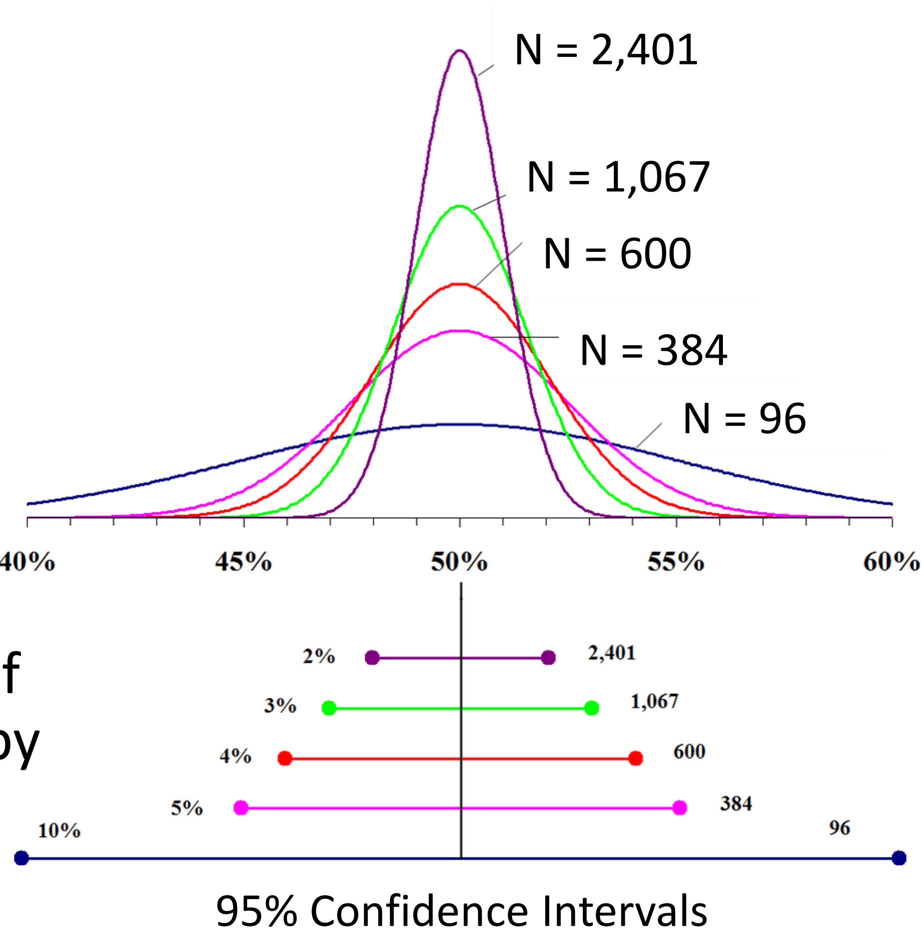
- A poorly designed study could generate very precise estimates that are completely wrong

Measures of uncertainty

We can see how uncertainty is reduced when more information is collected

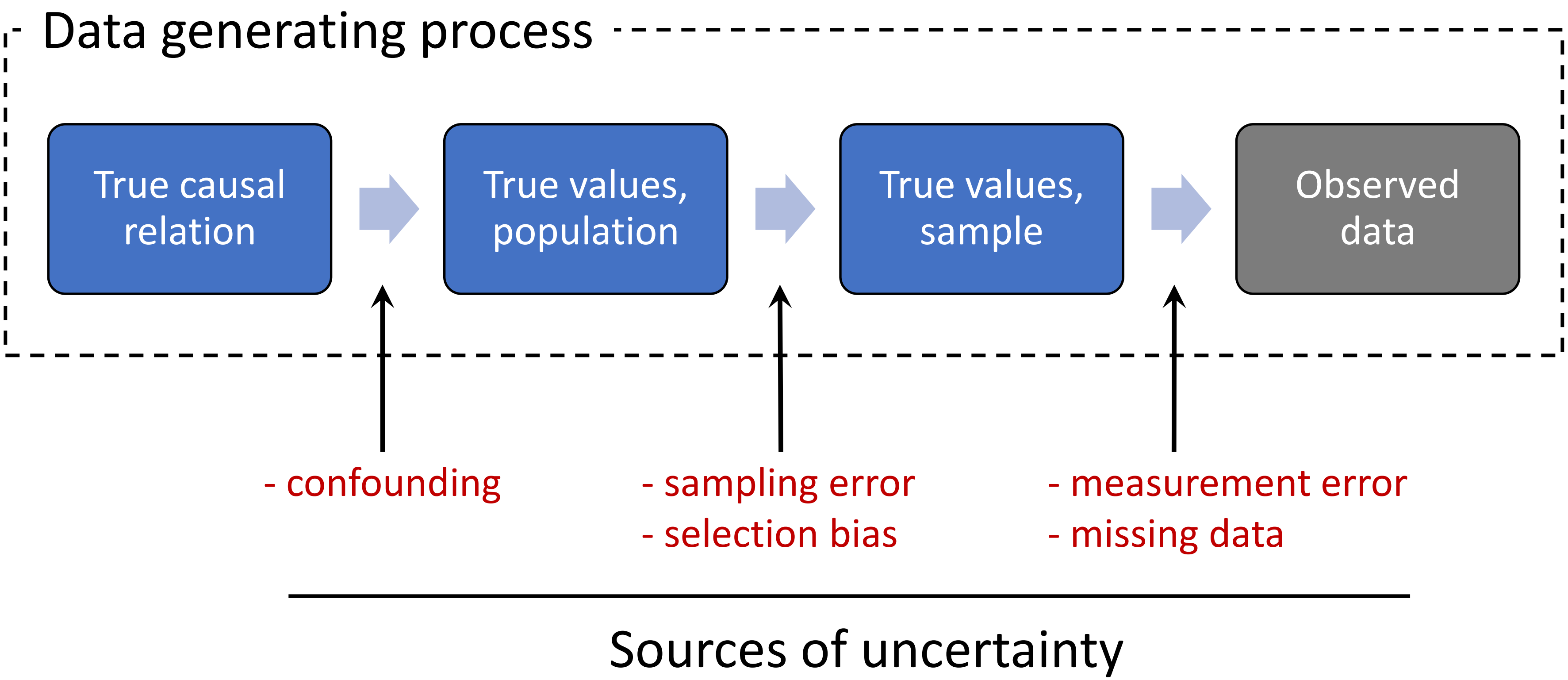
- Certainty that true (population) value is near observed (sample) value depends on sample size

- In order to reduce sampling uncertainty (95% CI) by a factor of 2, we must increase sample size by a factor of 4



https://en.wikipedia.org/wiki/Margin_of_error

Sources of uncertainty



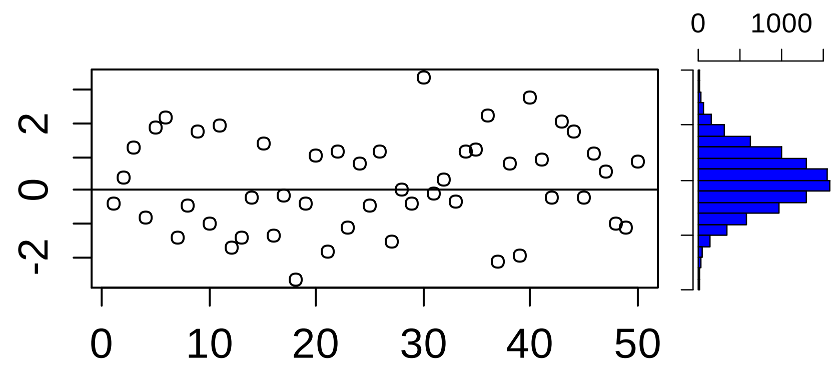
General linear models

- Recall that in **general linear models**, the error term (ε) captures so-called 'unexplained variation'

intercept predictor

$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

response slope error ('residual')



These errors (residuals) are assumed to be normally distributed

General linear models

- Early linear regressions were used to study positions of astronomical bodies and variation was due to measurement error, for which the Normal distribution is appropriate

- In later applications, particularly in biology, variation in measurements arose from both uncertainty and natural variability; hence, error term commonly includes any ‘unexplained variation’

Generalized linear models

- Recall that in **generalized linear models** (e.g., Poisson regression), no error term is specified

$$y_i \sim \text{Poisson}(u_i)$$
$$E(y_i|x_i) = log(\mu_i) = \alpha + \beta x_i$$

- But that the expected variance around our observations y comes directly from the underlying distribution
  - For example, in Poisson regression variance of y_i should be equal to the mean μ_i
- As with general linear models, the observed error can be a 'catch all' of uncertainty (randomness, measurement error) and natural variability

Summary

- Two related concepts, natural variability and uncertainty, generate measureable variation in our data
  - some of this variation is interesting (differences between individuals), but some is a nuisance that may or may not be reducible
- Samples are often used to study populations and this presents us with both natural variability and uncertainty
  - the two sources of variation have differing measures and interpretations
- In statistical modeling using GLMs or GLIMs, observed variation is usually captured in a ‘catch all’ of both natural variability and uncertainty
  - Knowing there are multiple sources of variation helps in the understanding of these models and what is actually explained