Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

Fundamentals of Statistics for Language Sciences LT2206

Jixing Li

Lecture 2: Sampling

Slides adapted from Cecilia Earls

Lecture plan

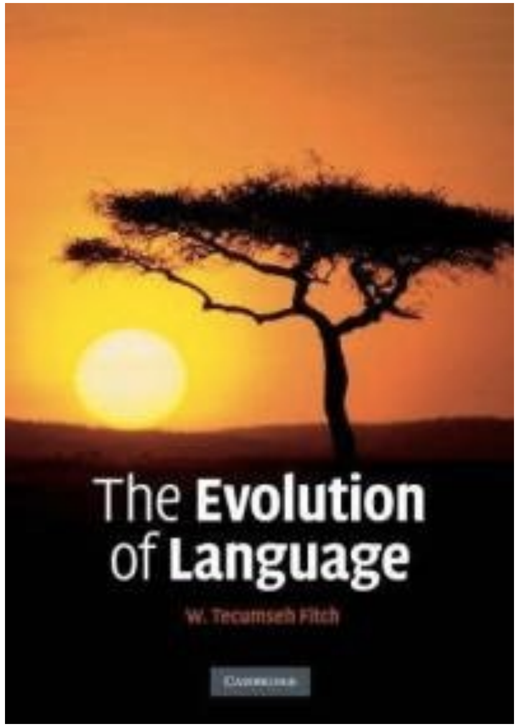
- Review on the three statistical problems
- Data sampling
- Short break (15 mins)
- Hands-on exercises

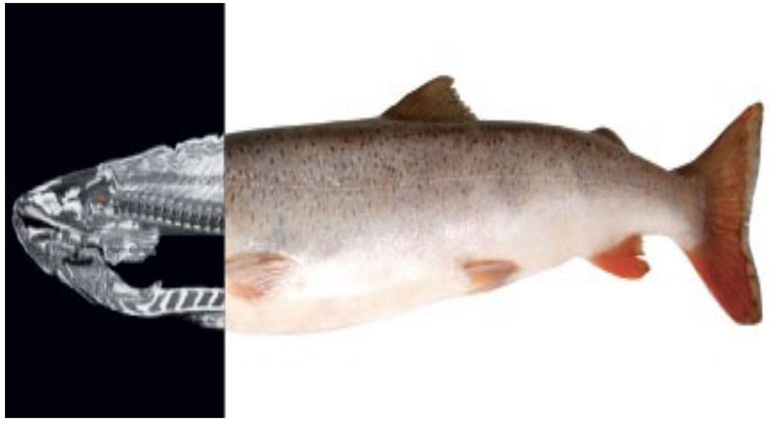
Three statistical problems

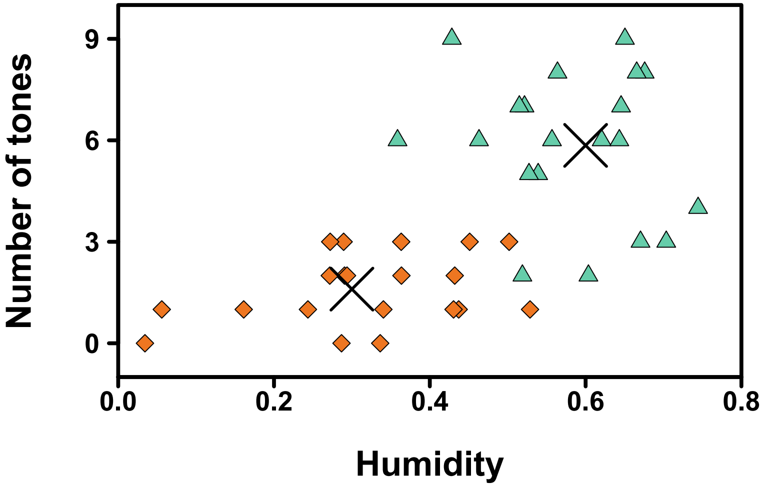
Correlation is not causation

Multiple comparisons

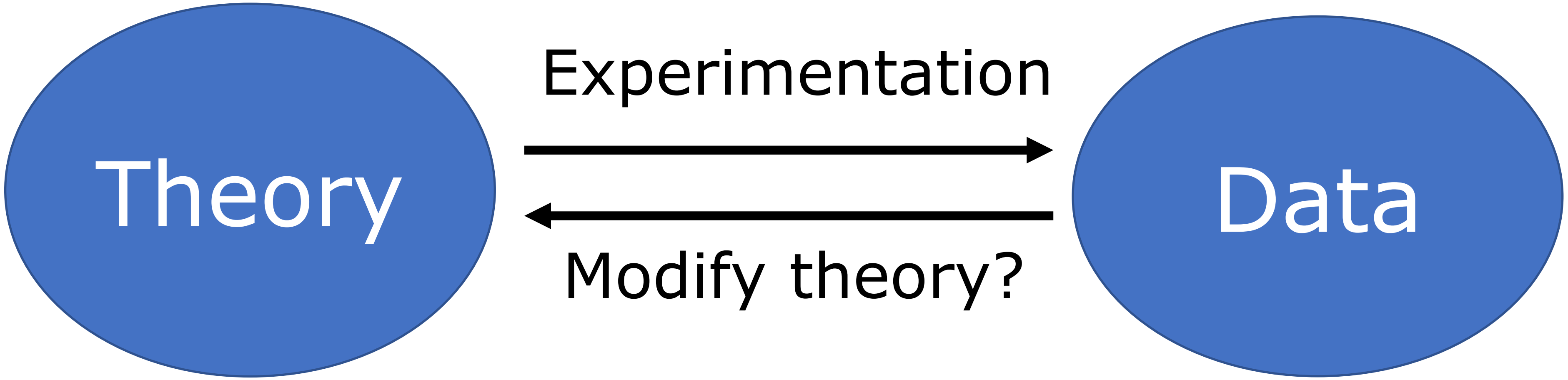
Lack of independence







Statistics: The science of learning from data

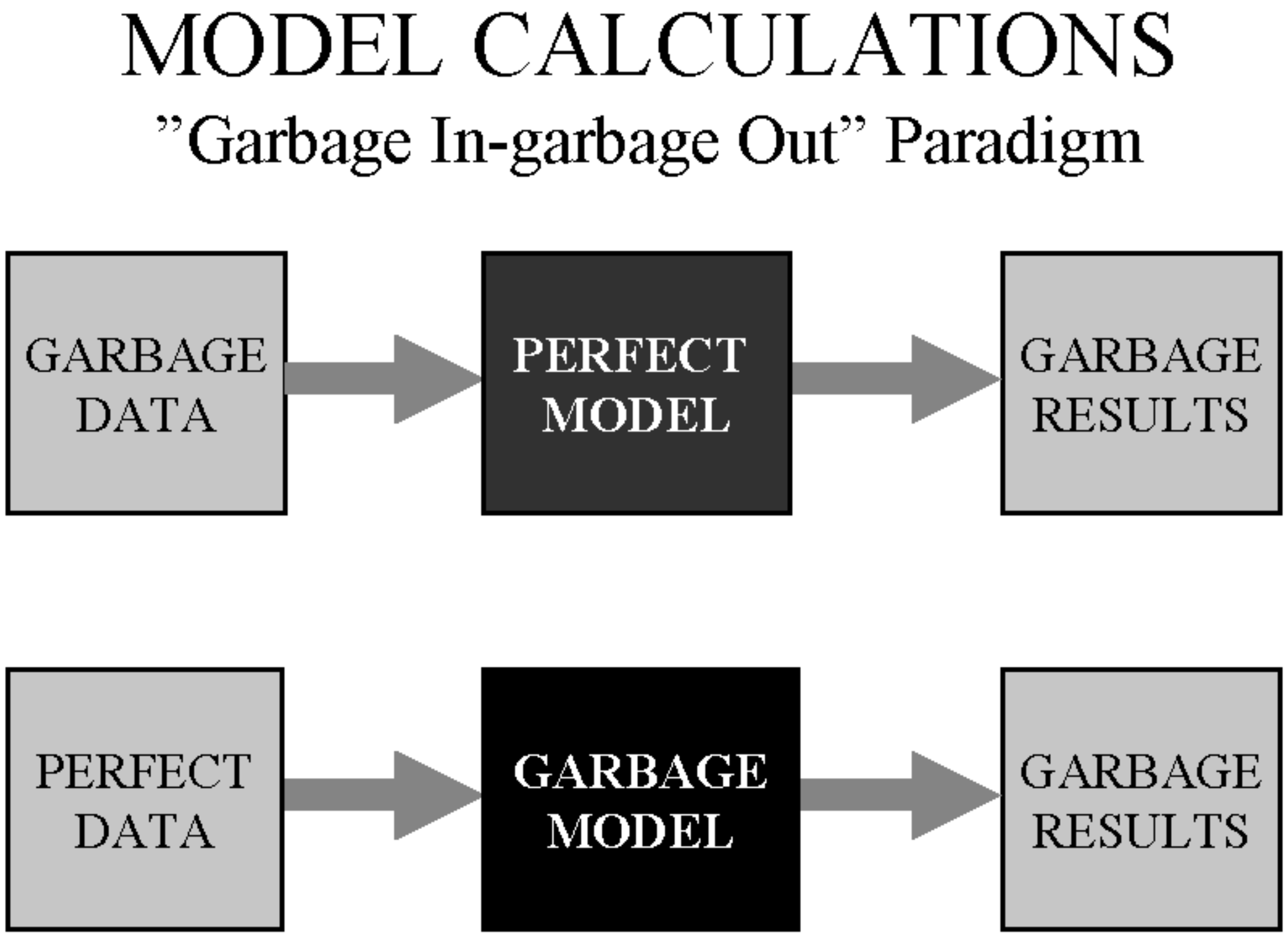


In general, statistics is concerned with:

- 1. Systematic methods for data collection
- 2. Objective methods for data analysis and "inference"
- 3. Careful interpretation of results

Science is a process for learning about nature in which competing ideas about how the world works are measured against observations (Feynman, 1965).

Two ways to assure your results are meaningless…

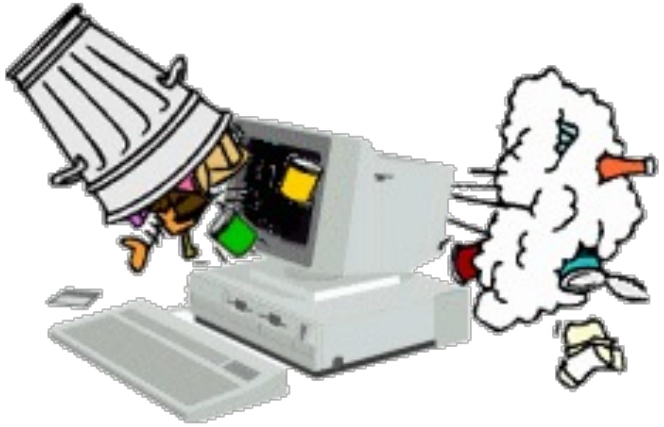


MODEL CALCULATIONS

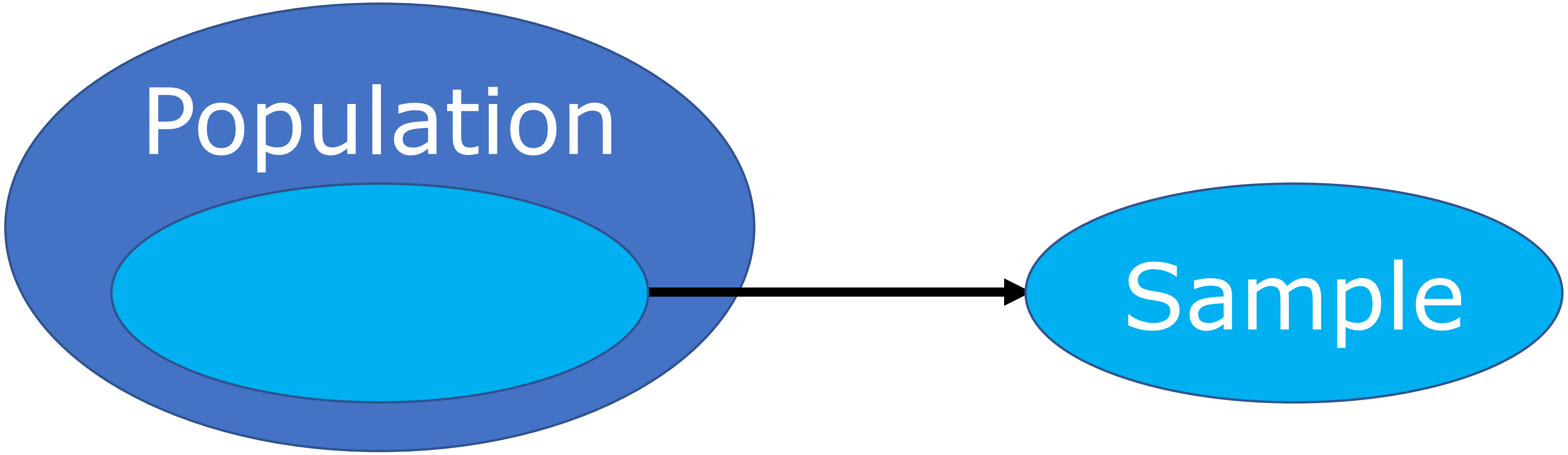
”Garbage In-garbage Out” Paradigm

GARBAGE DATA → PERFECT MODEL → GARBAGE RESULTS

PERFECT DATA → GARBAGE MODEL → GARBAGE RESULTS



Sampling: A crucial statistical concept



Population: the collection of all units of interest

Sample: any subset of units from the population

Unit/element/subject: individual entities that form the population when viewed as a group

Why should we sample? And how?

Why do we sample ?

"You don't have to eat the whole ox to know the meat is tough." -- Samuel Johnson

Examining the entire population may be:

- too expensive or slow to be feasible
- impractical if observation destroys the unit (rats, cars)

The goal of statistics is to learn about the population by examining only a fraction of it.

If a **small** fraction of the population gives an **accurate** picture of the population, we win big in speed and cost!

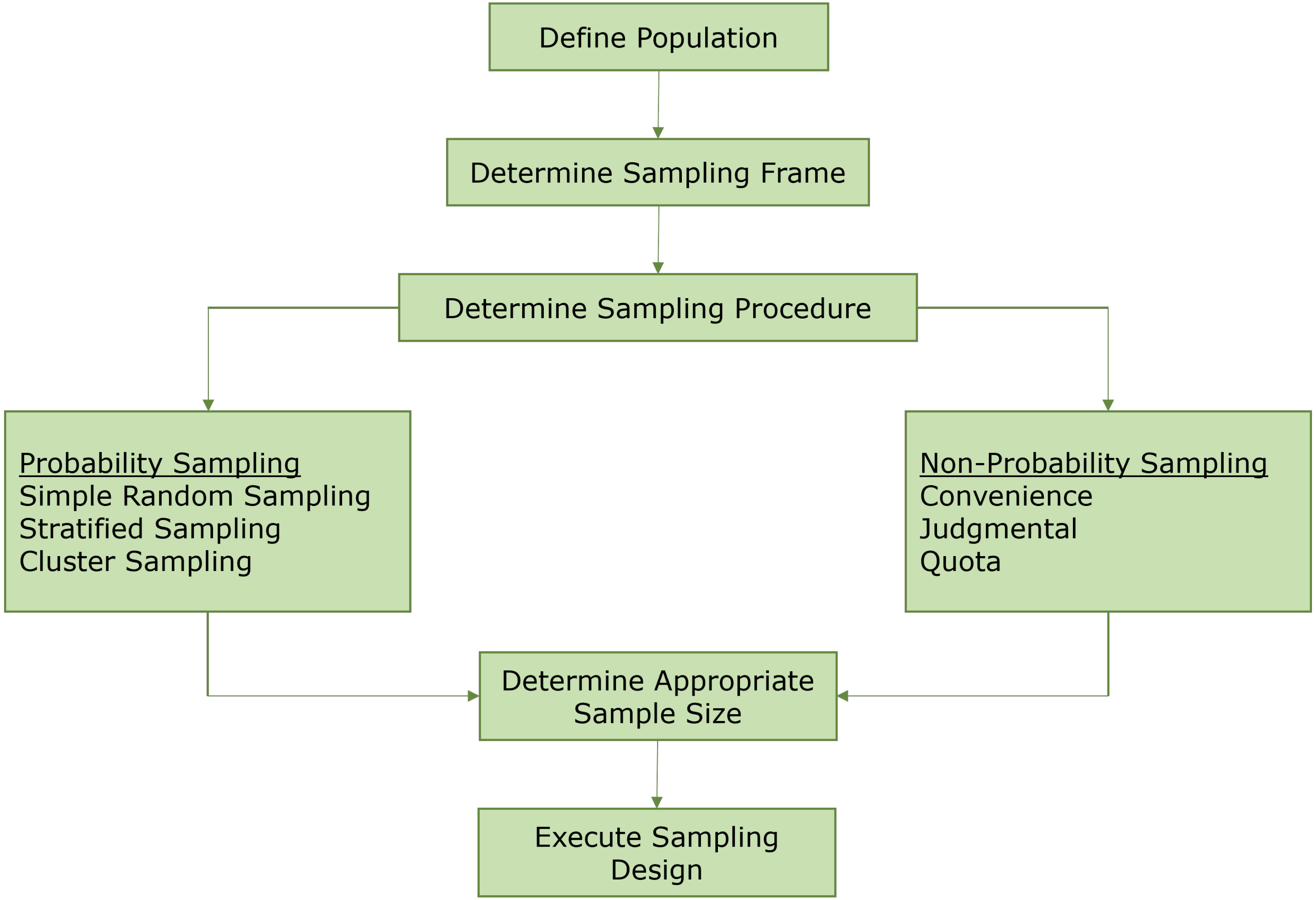
Evaluating sampling methods

Sampling methods can be good, bad, or terrible. What do these terms mean? They involve the validity and accuracy of conclusions from the sample.

Intuitively, a sampling method is good if it results (with high probability) in samples that are representative of the population.

How we sample depends on the objective of the study and what is practically feasible, but the goal is to use a good method.

Sampling design process



Define Population

Determine Sampling Frame

Determine Sampling Procedure

Probability Sampling
Simple Random Sampling
Stratified Sampling
Cluster Sampling

Non-Probability Sampling
Convenience
Judgmental
Quota

Determine Appropriate Sample Size

Execute Sampling Design

Target population

Addresses the question: Ideally, what collection of units would you like to describe? Be specific!

- Can you sample from the “ideal” population?
- Accessible population – the population you can sample from
- Accessible ≠ Target
- Significant sampling bias?
- Redefine scope?

Example: Student survey

The population of interest in such a study could be:

- All students in one class
- All students who attend CityU
- All students who attend a Hong Kong university
- All students in the world

The **unit:** an individual student

The **sample:** any subset of the population

Note: A sample that is good for one population is likely to be bad for a different population.

Determining the sampling frame

Enumerate the population; i.e., obtain a “list” of the population which will help you reach the sample.

Example: If the population of interest is all CityU students, we could consider using

- List from the registrar
- Phone book
- Student union listing
- University mailing list

Problems with lists: access, omissions, out-of-date, duplicates

Selecting a sampling design

Probability sampling –any sampling method based on a random selection process

- simple random sampling –the Gold standard
- systematic sampling
- stratified sampling
- cluster sampling

Non-probability sampling –non-random sampling

- convenience sampling
- judgment sampling
- snowball sampling (response-driven sampling)
- quota sampling

Simple random sampling

A sample of size *n* is a **simple random sample (SRS),** if it is selected by a method that gives every possible sample of *n* units the same probability of being chosen.

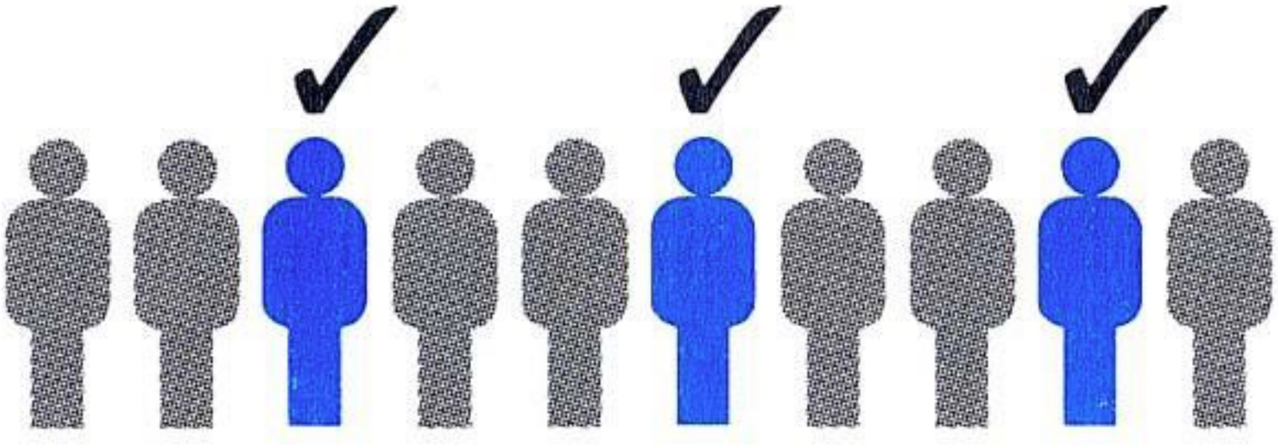
We could think about how to obtain an SRS of all students in a class or a university

Are the students in a class who respond to the student survey an SRS of…

- all students in the class?
- all students at CityU?

Systematic sampling

- Order all units in the sampling frame based on some variable and number them from 1 to *n*.
- Choose a random starting place from 1 to *k* and then sample every *k*th unit.
- The choice of *k* depends on *n*.



Stratified sampling

The chosen sample is forced to contain units from each of the segments, or strata, of the population.

Goal: equalize "important" variables; e.g., gender, race, school, geographical area, etc.

Procedure:

- Divide population into mutually exclusive and exhaustive strata based on an appropriate population characteristic (the “important variables”).
- Draw simple random samples from each stratum.

Stratified sampling

Proportionate stratified random sampling: Sample size from each stratum reflects the proportion of the population that belongs to the stratum

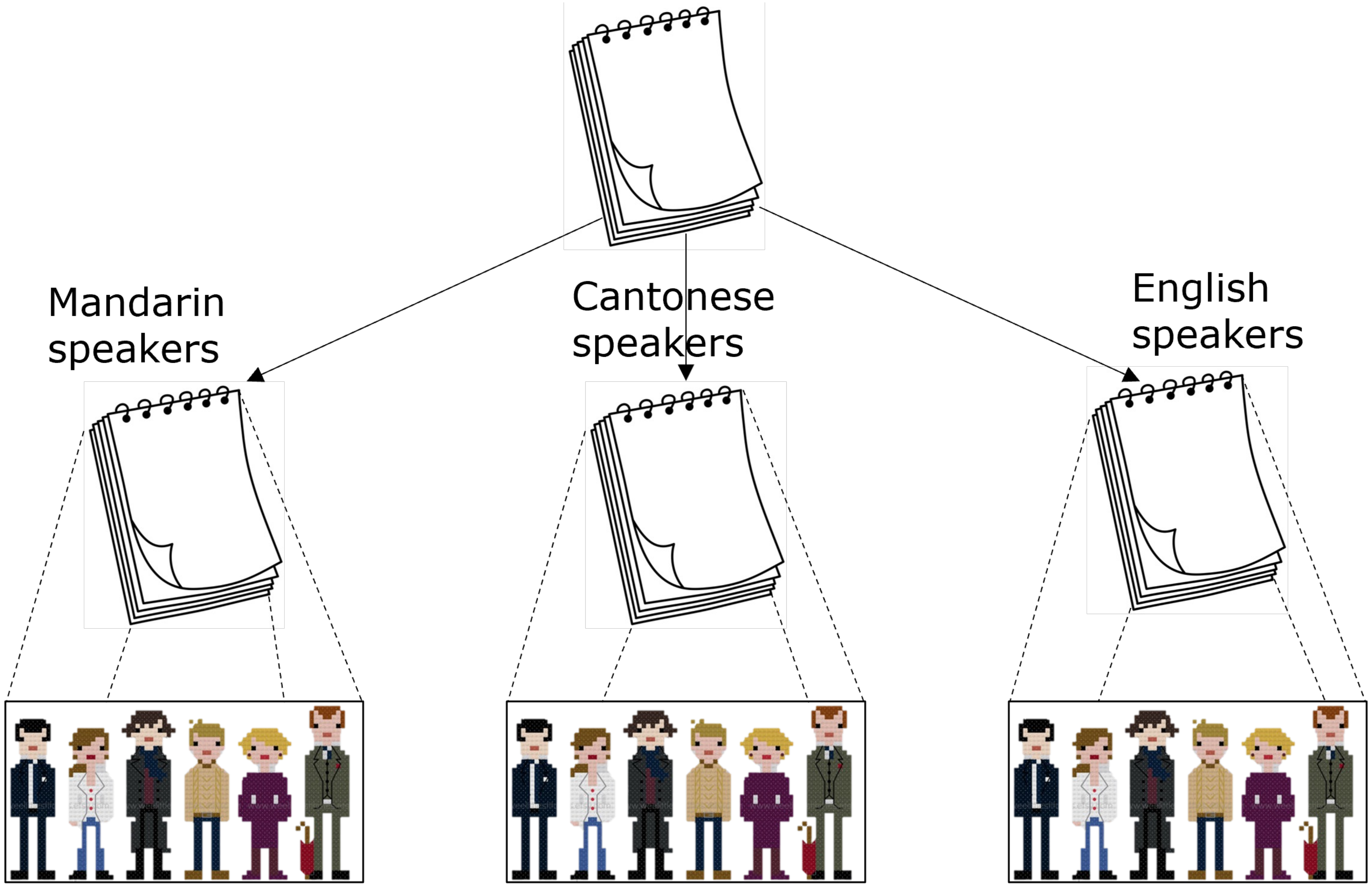
Advantages

- Smaller sampling error than simple random sample: one source of variation is eliminated
- Usually ensures representativeness

Disadvantage

- May fail to represent the target population

Stratified sampling



Cluster sampling

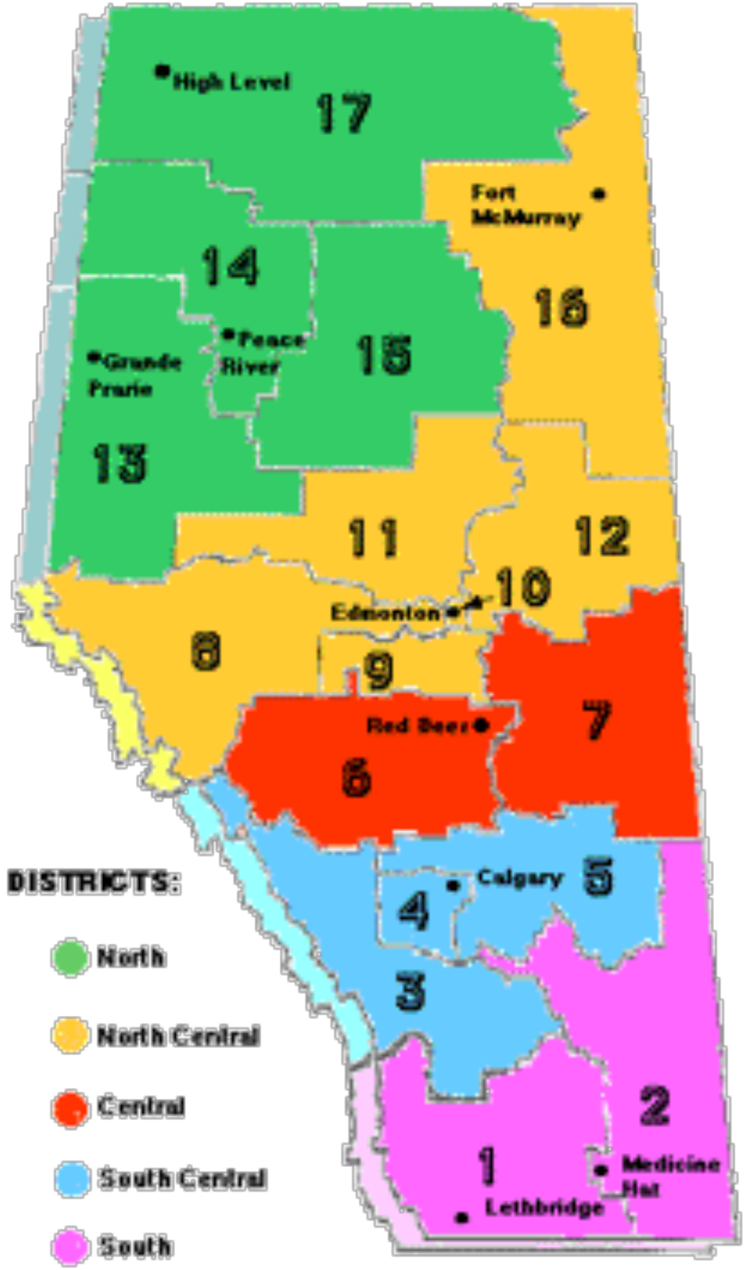
Clusters of population units are selected at random and then all or some randomly chosen units in the selected clusters are sampled.

Procedure:

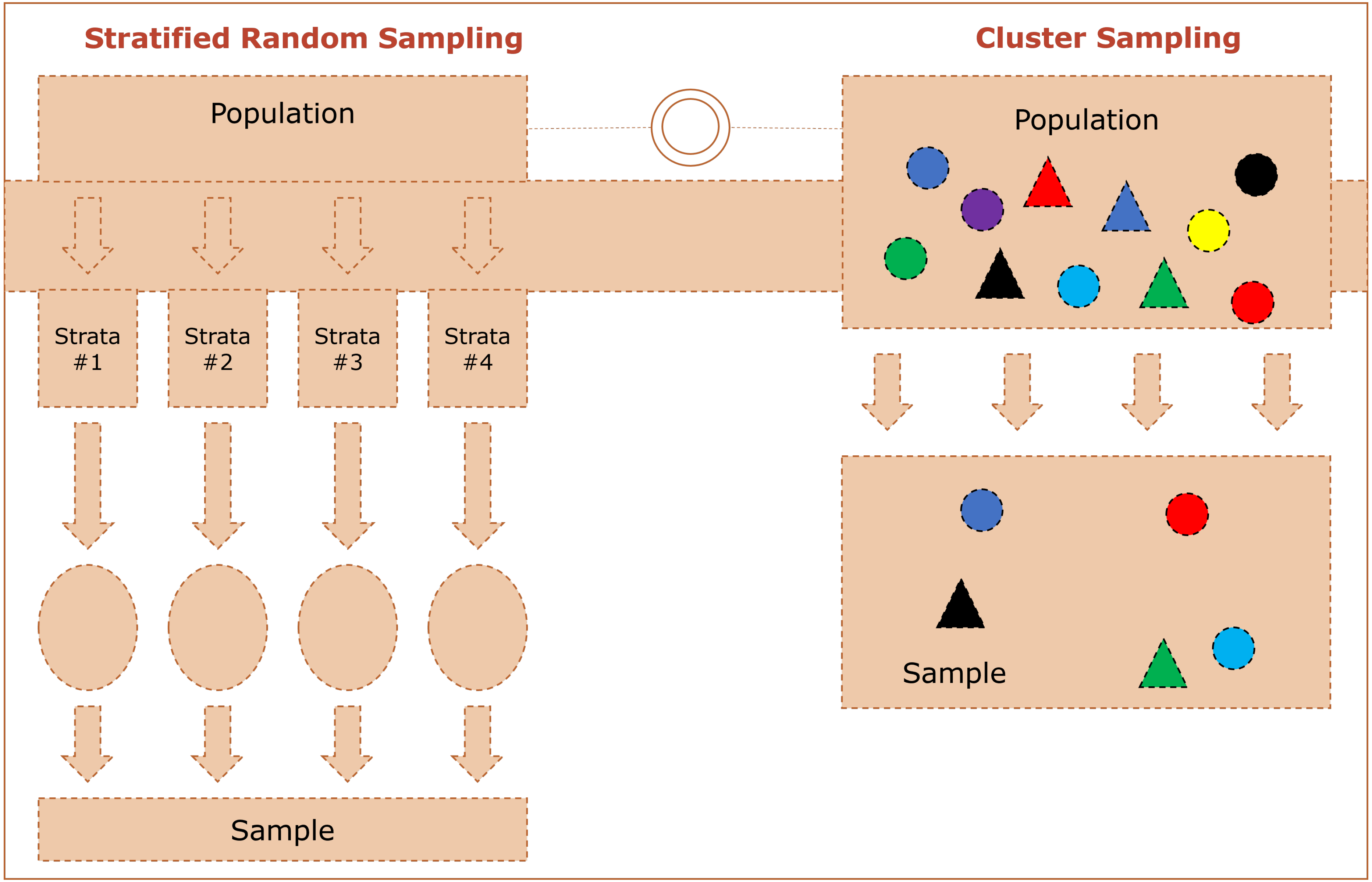
- Population is divided into mutually exclusive and exhaustive subgroups, or clusters. Ideally, each cluster adequately represents the population.
- A simple random sample of a few clusters is selected.
- All or some randomly chosen units in the selected clusters are studied.

Cluster sampling by region

- Divide population into clusters, usually along geographic boundaries.
- Randomly sample clusters.
- Measure units within sampled clusters.



Stratified vs. cluster sampling



Selecting a sampling design

Use stratified sampling when:

- The primary research objective is to compare groups.
- Using stratified sampling may reduce sampling errors.

Use cluster sampling when:

- There are substantial fixed costs associated with each data collection location.
- When there is a list of clusters available but not of individual population members.

Non-probability sampling

Subjective procedure in which the probability of selection for some population units is zero or is unknown before drawing the sample.

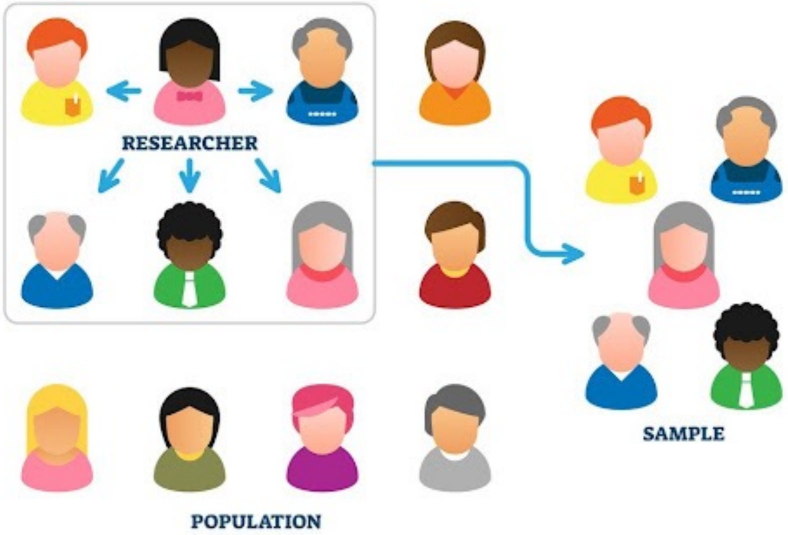
- Information is obtained from a non-representative sample of the population.
- Sampling error can not be computed.
- Results cannot be projected to the population.
- Cheaper and faster than probability sampling, but gives limited inference.

Types of non-probability sampling

Convenience sampling

A researcher's convenience forms the basis for selecting a sample.

- Student volunteers (e.g., undergraduate linguistics students)
- "Man on the street" interviews



Judgment sampling

A researcher exerts some effort in selecting a sample that seems to be most appropriate for the study.

Types of non-probability sampling

Snowball sampling

The selection of additional respondents is based on referrals from the initial respondents.

- Referral sampling: Friends of friends

Used to sample from low-incidence or rare populations

Quota sampling

The population is divided into cells on the basis of relevant control characteristics (similar to stratified sampling).

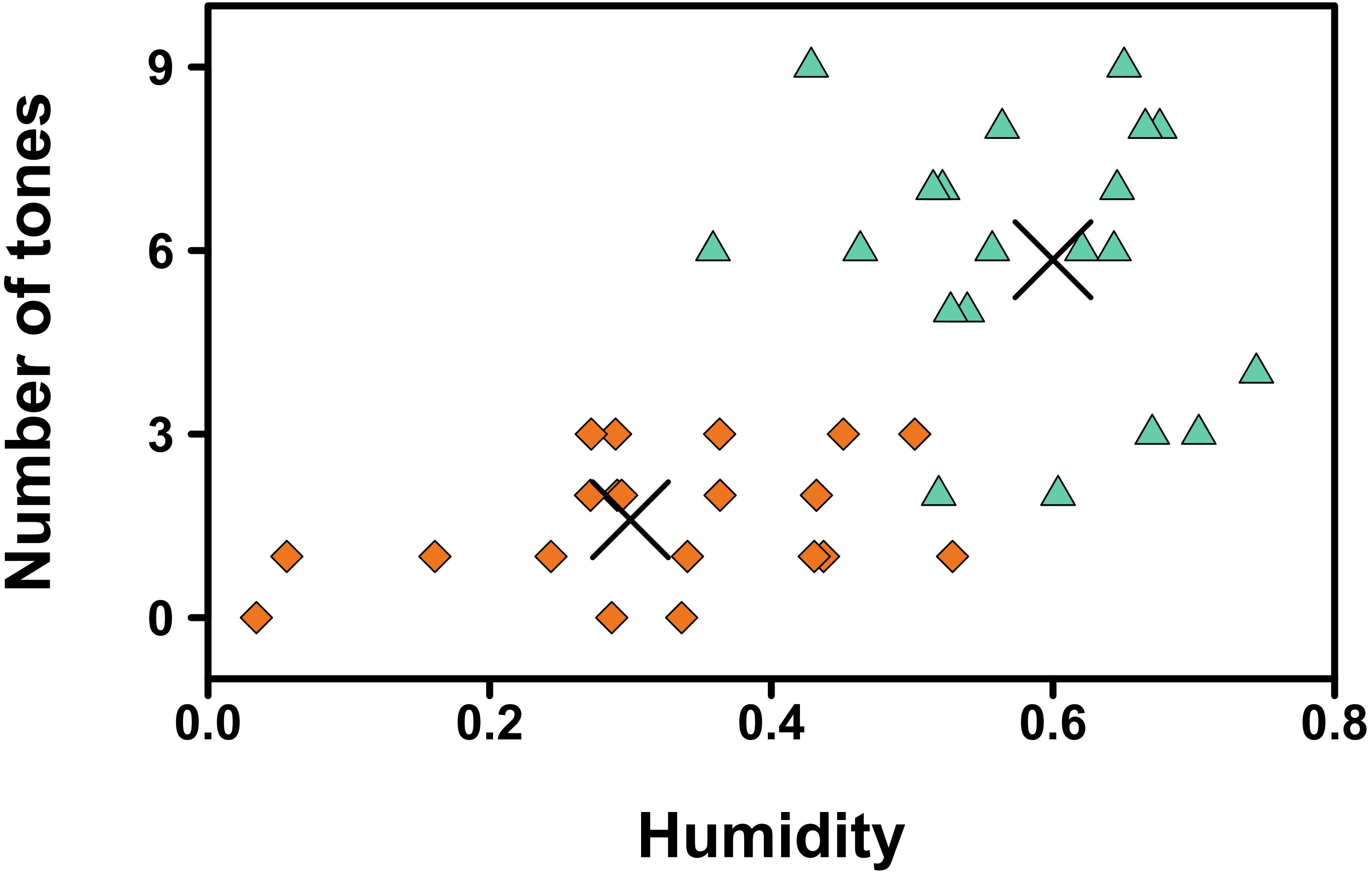
- A quota of sample units is established for each cell (e.g., 50 women, 50 men)
- A convenience (or judgment) sample is drawn from each cell until the quota is met.

Probability vs. Non-probability sampling

- Non-probability sampling may be less time-consuming and less expensive.
- May give you some idea about population characteristics, but nothing can be said with any certainty.
- Quantitative generalizations about the population can only be done under probability sampling.
- Drawing inference from samples with a non-representative sample is dangerous

  e.g. television news show’s viewer polls

Sampling error



Errors in sampling

Random sampling error: The sample selected is not representative of the population due to chance.

- The amount of random error is controlled by sample size; a larger sample size usually leads to a smaller sampling error.

Non-sampling error: Systematic error, not controlled by sample size.

- **Non-response error:** units selected in the sample do not respond in whole; only an issue if non-responders are different than those that did respond
- **Respondent error** (e.g., lying, forgetting, etc.)
  - Interviewer bias
  - Recording errors
  - Poorly designed questionnaires

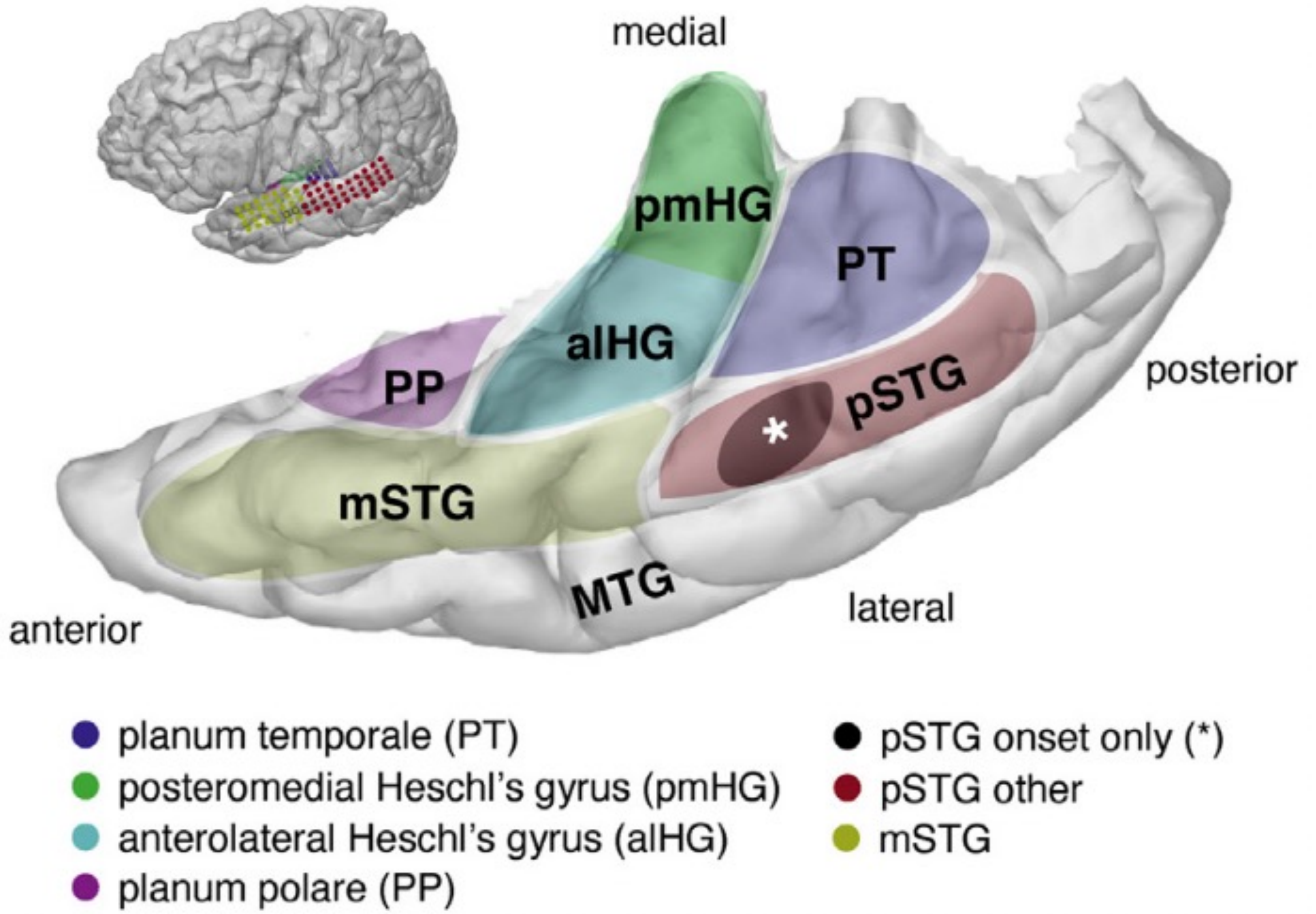
Sampling limitations

- **Uncertainty due to sampling:** Measurement (almost) always involves uncertainty due to the need to use a sample rather than the whole population.

- **Other sources of error:** Inaccurate or incomplete recording of data; i.e., selective sampling of data. These add to the uncertainty and can adversely affect results and/or interpretation. Counteracting these effects is usually difficult and methods typically depend on strong, untestable assumptions.

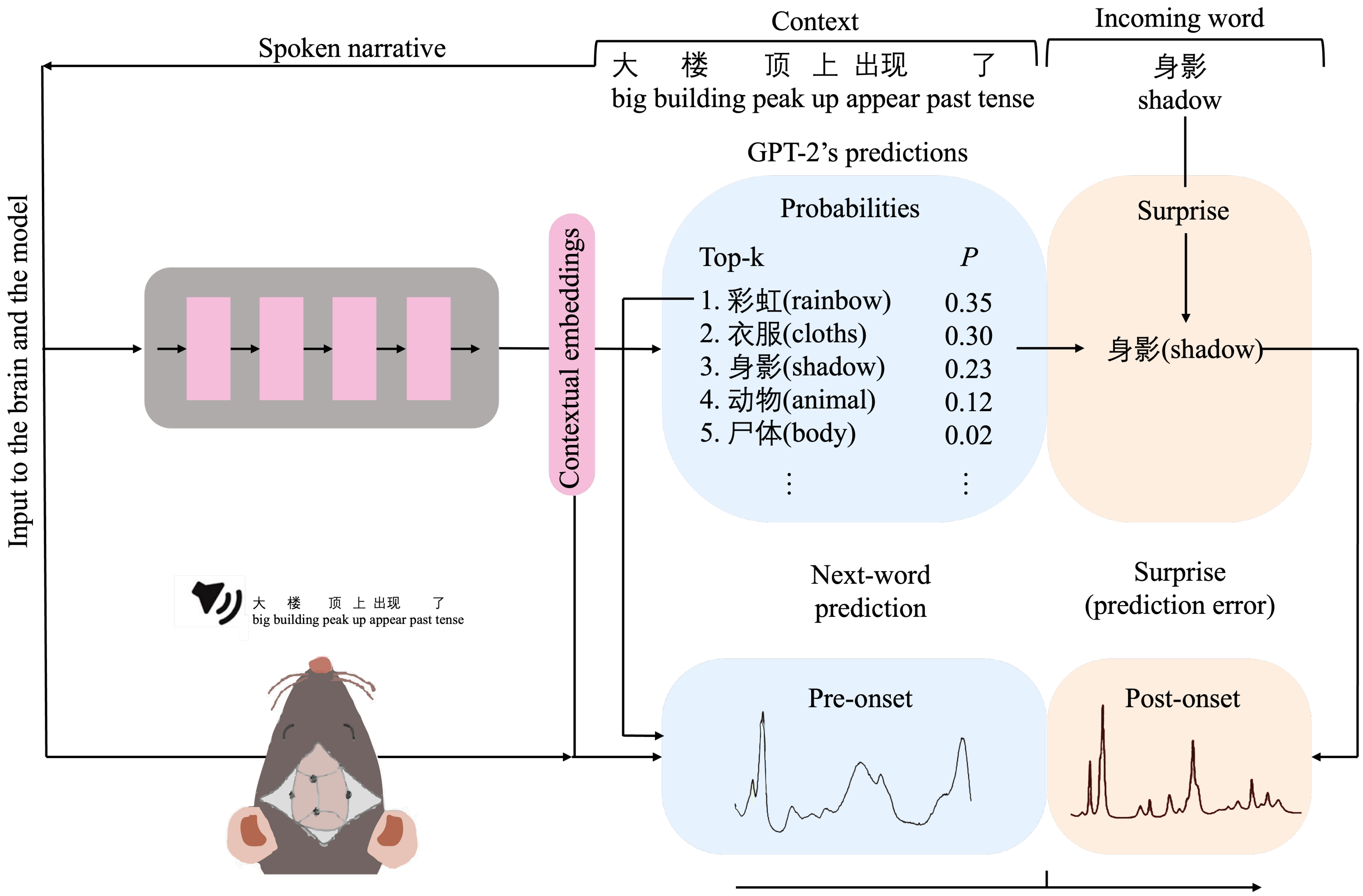
Example: ECoG experiment

Example: ECoG data



Hamilton et al. (2021)

Example: Rat data



Spoken narrative

Context

大 楼 顶 上 出现 了

big building peak up appear past tense

Incoming word

身影

shadow

Input to the brain and the model

Contextual embeddings

GPT-2's predictions

Probabilities

Top-k	*P*
1. 彩虹(rainbow)	0.35
2. 衣服(cloths)	0.30
3. 身影(shadow)	0.23
4. 动物(animal)	0.12
5. 尸体(body)	0.02
⋮	⋮

Surprise

身影(shadow)

大 楼 顶 上 出现 了

big building peak up appear past tense

Next-word prediction

Surprise (prediction error)

Pre-onset

Post-onset

To do

- Install R and R Studio on your laptop
- Check out Lab 2
- Read: Next lecture: Textbook Ch3