Frontiers of Statistics

Emerging applications

- Computer vision
- Recommender systems
- Predictive analytics
- Fraud and anomaly detection
- Risk assessment
- Social and government services

Where Do Data Come From?

Brady T. West

Different Types of Data

- **Two key types of data:**
  - Organic / Process Data
  - "Designed" Data Collection

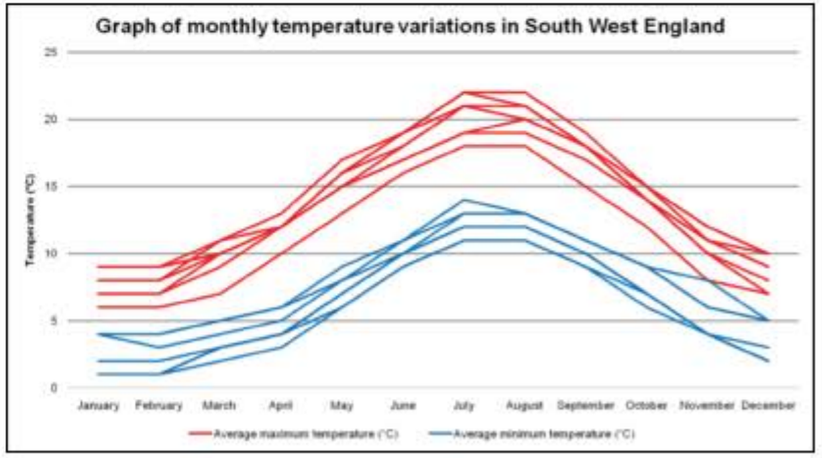
Organic / Process Data

- Generated by computerized information system, or extracted from video / audio recordings
- Generated "organically" as the result of some process, often over time

Organic / Process Data

Examples

- Financial or Point-of-sale transactions/ Stock market exchanges
- Netflix viewing history
- Web browser activity
- Sporting events
- **Temperature/pollution sensors**



Organic / Process Data

- These processes generate massive quantities of data

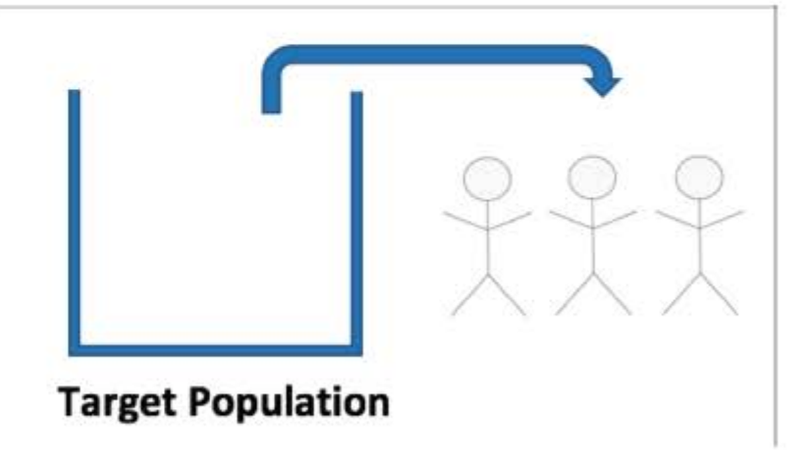
→**"Big Data"**

- E.g., baseball games, meals ordered at McDonald's on a given day, changes in temperature on a given day in a particular city
- Processing requires significant computational resources; **data scientists** "**mine**" these data to study trends and uncover interesting relationships

"Designed" Data Collection

Designed to specifically address a stated research objective

- Individuals sampled from a population, interviewed about opinions on a particular topic
- Tweets extracted from Twitter and coded to analyze how often people are expressing opinions about a particular topic



"Designed" Data Collection

Common features of "designed" data

- Sampling from populations, administration of carefully designed questions
- Typically data sets much smaller compared to organic/process data sets
- Data collected for very specific reasons, rather than simple reflections of ongoing natural process

Will work with both types ~ more on sampling later in this course

Are the Data i.i.d.?

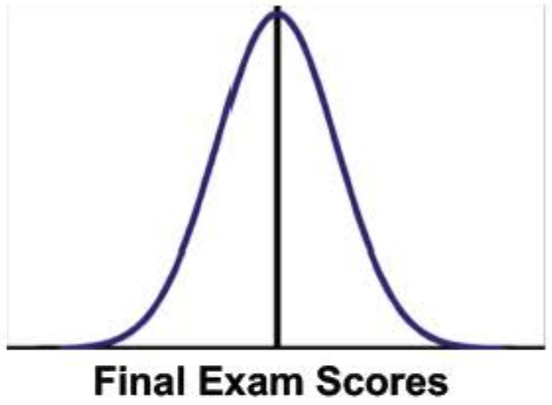
For analyzing data, regardless of source, an important question:

Q: Can the Data be considered **i.i.d.?**

i = independent and id = identically distributed

Observations on variable of interest are completely independent of all other observations (no correlation!) and arising from a common distribution

i.i.d. Data



- **Example: Final exam scores** from a large Intro to Stats class at a university are **independent observations** from a **common normal distribution**
- **Can estimate features** of that distribution (mean, variance, extreme percentiles), and **make inference** about those features with a certain amount of precision

What if Data are NOT i.i.d.?

Examples

- Students sitting next to each other tend to have similar scores
- Males and females might have different means
- Students from same discussion section may have similar scores

Dependencies and differences need to be accounted for in analysis!

→ **Need different analytic procedures**

Important Notes

- Need to Ask:
  Can we can apply procedures that **assume i.i.d. data**?
- Always **consider where data came from**!

Later in this course:

More on "designed" data collection and the i.i.d. idea!