Trace evidence and archival data

Updates

- Review materials posted soon
- Old quiz questions posted soon
- Final exam: 2/3 new, 1/3 old
  - Old review materials still apply
  - Similar format: some fill-ins, some shortish-answer
- Review session **Monday March 16** (time, place TBA)

- Next time (Wed): bring a computer or smartphone to do evaluations (end of class)

Updates Part II

- “Quiz” this week: posttest
- 60 minutes to complete
- You can look over your notes and book before/during, but you don’t have to
- Graded on **legitimate attempt to complete**
  - Not graded on accuracy
  - So if you take this seriously you will get a 100% for one of your quiz grades
- After deadline closes, it will show you your answer and the correct answer as preparation for the test

Two parts

- Physical evidence
- Archival evidence
- What’s nice about both of these:
  - Evidence is __________ (doesn’t change _________)

**Big Data**
Physical trace studies

• Study _______ evidence that results from _________
• ________ (by-product of behavior)
• ________ (people purposely create)
• ________ (buildup)
• ________ (wearing of material)
• ________ trace measure (no researcher ______ needed)
• ________ trace measure (researcher sets up situation to _________ behavior)

Physical trace studies: cautions

• Selective _________
  – Some materials wear more rapidly than others
• Selective _________
  – Not all individuals will leave the same traces
  – (Litter, _________, etc.)
• ________ issues
  – Traces may reveal _________ information
  – Risks to _________ collecting evidence

Archival data

• Written or _________-stored _________
  – Tax returns
  – Videos
  – _________, _________
• ________ (running) records
  – Maintained on a regular basis
• _________ records
  – Happen sporadically or _________

Physical trace studies

Professor Q wants to know what types of foods are eaten by the mountain coati, a very secretive raccoon-like animal. They have rarely been sighted and do not respond to desensitization. To gauge their eating habits, Professor Q searches the Venezuelan forests for their old sleeping nests and looks for food debris.

Archival data

• Written or _________-stored _________
  – Tax returns
  – Videos
  – _________, _________
• ________ (running) records
  – Maintained on a regular basis
• _________ records
  – Happen sporadically or _________
Archival data

• Records
  — Created explicitly for other people to view/hear
• ________
  — Created more for personal use
• Ambiguous: “This is the worst day of my life!”
  — Private blog/diary: a ________
  — ________ post: a record
• Often, too much data: need a ________
  — Data reduction (like ________ studies)
  — ________ analysis (d.r. specifically for written records)

Archival data concerns

• Selective ________ (similar to selection ________)
  — Certain people may be more likely to make ________
  — Problem: records are not ________ of the population under study
  — Editorial policy of (e) ________ (marriage announcements, birth records, letters to editor)
• Selective ________
  — Friendster, anyone?
  — Document shredding
  — Documents/evidence destroyed in ________
• Ethics
  — For small subsets of data, may reveal ________ information
  — ________ datasets may violate privacy (members of a religious group vs. purchases made at ________)

Archival data

Professor R wants to examine the emergence of the “pink princess” phenomenon—the tendency for young American female children to prefer to wear the color pink. R goes to a local dance school and finds photographs of little girls’ ballet recitals dating from 1950; she also finds photographs from a local church Easter picnic from around 1952. She notes the color of each young girl’s costume/dress and the year the picture was taken.

What is “big data”?

• According to SAS (a business analytics firm):¹
  • Four V’s and a C:
    — Volume: only now do we have enough storage capacity
    — ________: data comes in rapidly
    — Variety: lots of different types of files/information
    — Variability: rate of data can ________ (e.g. deadlines)
    — ________: need to ________ various data inputs to understand it in context


What kind of data?
- ________ (on articles, for-sale items), purchases
- Text: corpus (pl ________) – news articles, transcribed___________
  - English ________ Corpus
  - Google Books (books.google.com/ngrams)
  - Facebook posts, ________
  - All TOEFL essays for the past 5 years
  - May be raw or ________
- Image sets (faces, objects, ________)________
- ________ sequences
  - UK: 100,000 genomes project
  - http://www.genomicsengland.co.uk/the-100000-genomes-project/
- Basically anything you can imagine

Volume
- Can’t just have ________ code the data
- May select a ________ and have it coded
- May ________ analyze a lot of the data and spot-check it for ________

What is it good for?
- Examining ________
- Field experiments
- ________ analyses
- Verification of your new ________ ________
- ________ association studies (GWAS)
  - http://www.genome.gov/20019523

Big data, big cautions
- Selective ________: who creates the data?
  - Journalists (speech patterns)
  - Disgruntled commenters, ________
  - People who don’t mind ________
  - People who are interested in “brain training”
  - People who have concerns about ________ issues
  - ________ people (____ ____ _______ ____)
- Are big data ________?
- ________ survival
  - Mostly ________ genomes
  - No recordings of _____________ from 400 y ago
Big data

Professor S, a clinical psychologist, wants to know whether, among people aged 20-29, women are generally more depressed than men. To ask this question, S obtains Facebook posts from 1,000,000 individuals in the target age range. (Data are anonymized, with no names or location information, just age and gender.) S searches the text for depression-related words for women and for men.