Sampling and Statistical Significance

Main goal: Students understand some pitfalls and methods of sampling, and the basic ideas behind assessing statistical significance when using sampled data

1. When and why sampling occurs
2. Pitfalls
3. Sampling methods
4. Drawing conclusions from sampled data
    - “Noise” in sampled data
    - Using sampled data to support hypotheses
    - Using sampled data to make predictions

When and why does sampling occur?

1. If stored/streamed data is too large to process all of it, take a subset. Ex: tweets, images of sky at high granularity per second, every stock trade

2. Data reflects only a portion of actual events. Ex: data from selected people, sensors without full coverage in space and/or time, repeated trials of nondeterministic nature (e.g., network speed)

Sampling pitfalls

Sampling bias
- Landon vs. Roosevelt
- Hurricane Sandy tweets
- Those who respond to a poll
- Sensors that didn’t fail
- iPhones vs Androids

Confirmation bias
- “Seeking information that reinforces one’s beliefs” (type of cognitive bias), can also creep into sampling
- Ex: Water quality sensors all far from sewage outflow; create poll on iPhones only

Sampling methods

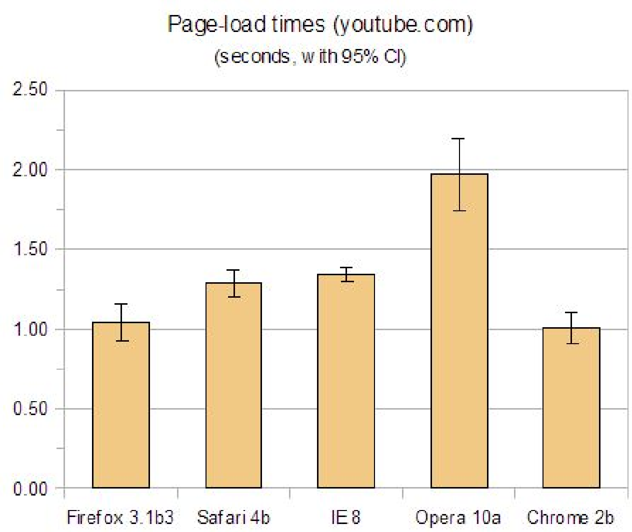
Goal: Sample should be accurate representation of the whole
Many methods, best one depends on setting. Ex:
- Simple random sampling
- Round-robin (systematic) sampling -- every Nth item from unordered set
- Stratified sampling -- proportionate random sampling from subgroups
- Judgement sampling -- choose representative subgroup

<u>Drawing conclusions from sampled data</u>

Measuring "noise" in sampled data

Confidence interval (type of "error bar")

- Compute an expected or average value *V* from a sampled set of values; confidence interval gives range around *V* that new value would fall in with X% probability
- Ex: 1000 network speed measurements, average 27.342, 95% confidence interval [24.5,31.23]
- Ex:



Using sampled data to support hypotheses

P-value

- Measures how likely it is that results achieved were an accident
- Conversely: If a different sample is used, what's the chance of getting a different result?
- Social scientists aim for $P < 0.05$, hard scientists aim for $P < 0.01$
- Ex: 1000 network speed measurements, claim average network speed is under 30. P-value = .03 says 97% chance if we ran the experiment again we'd still be under 30.
- Example: Upgrade network, claim average speed is at least 25% better. Took 1000 measurements before upgrade and 1000 measurements after upgrade showing 25% improvement. P-value = .06 says 94% chance if we ran the experiment again we'd still be at least 25% better.
- P-value too high: try increasing sample size

Using sampled data to make predictions

"Goodness of fit"

- How closely a model (function) used for prediction matches the sample
- Ex: linear regression, R-squared measure, the smaller the better