Squeezing the juice out of linguistic data:
Statistics in Linguistics

Mike Hammond

Ying Lin

Natasha Warner

Andy Wedel

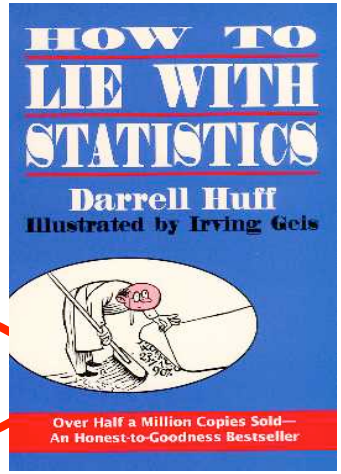
Discussant: Sandiway Fong

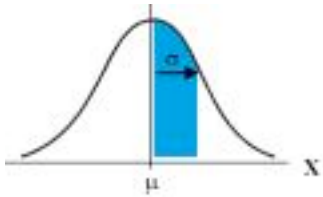
Go Study Statistics!

- from yesterday's *New York Times*:

Take a course in neuroscience. In the next 50 years, half the explanations you hear for human behavior are going to involve brain structure and function. You've got to know which are serious and which are cockamamie.

Take statistics. Sorry, but you'll find later in life that it's handy to know what a standard deviation is.

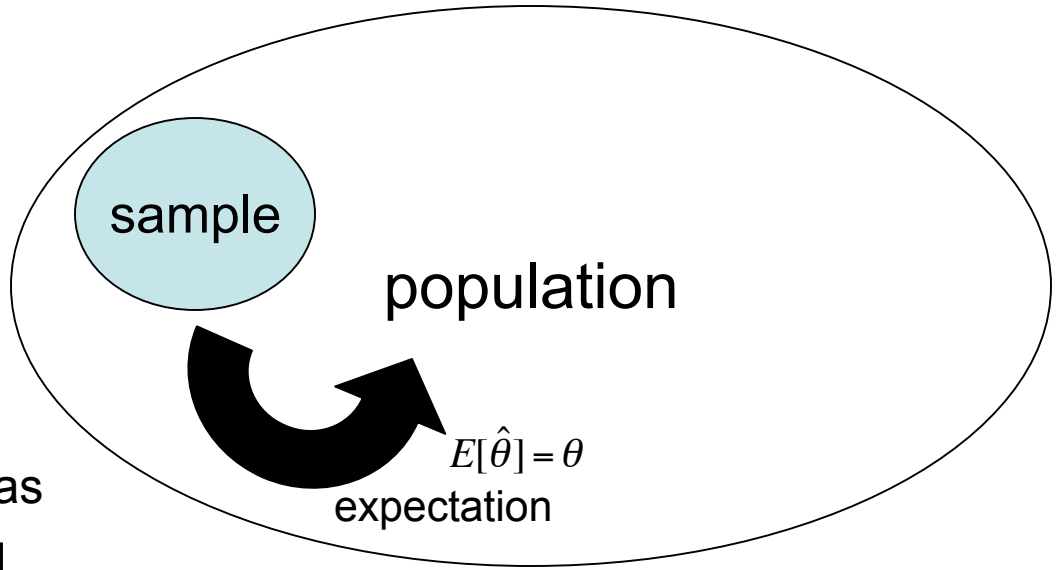




Big Picture

- want results about a sample to generalize to the population

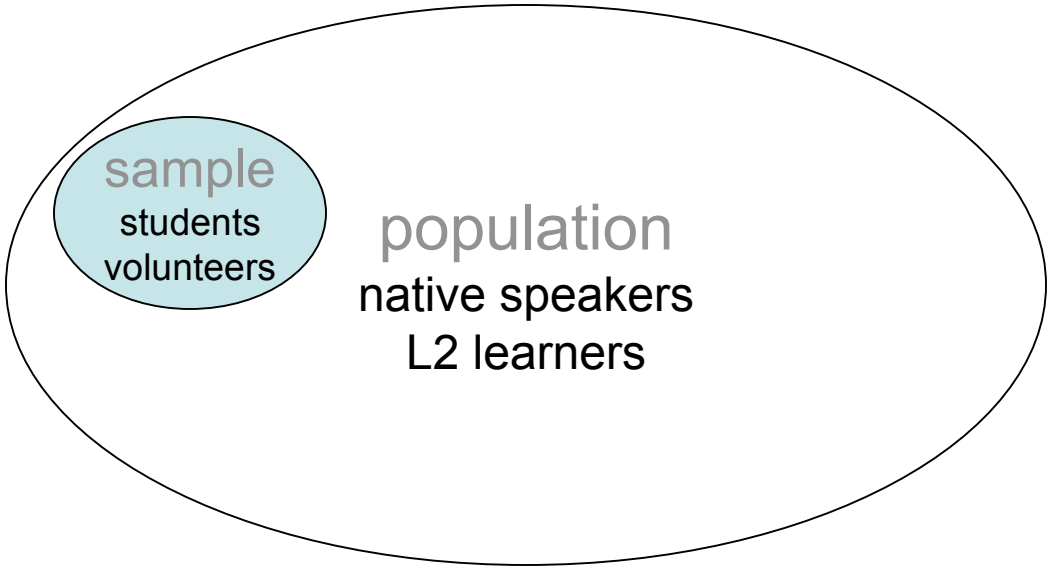
only governments can afford to take a census



sometimes
referred to as
Learning

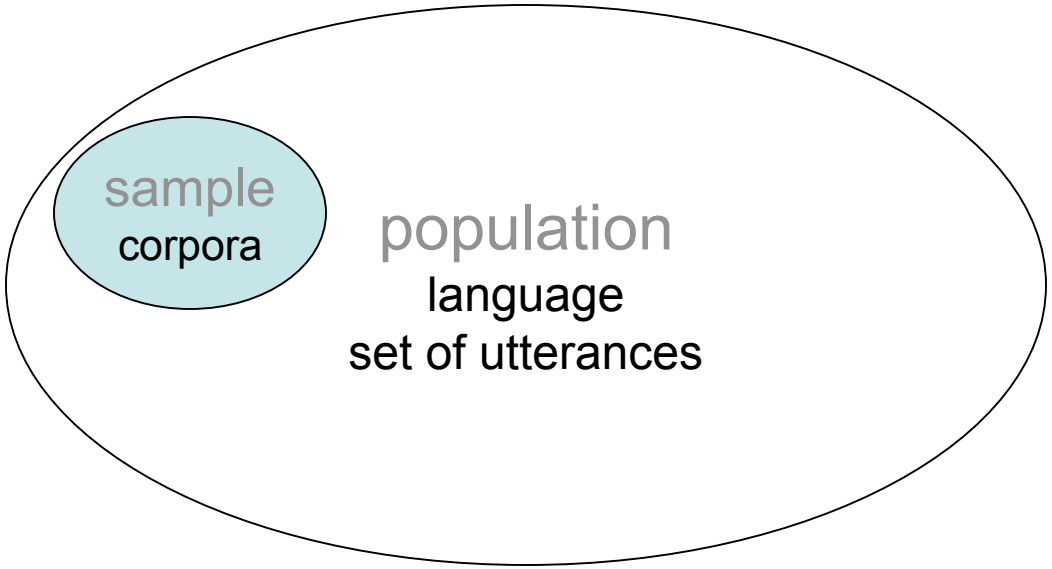
Big Picture

- want results about a sample to generalize to the population



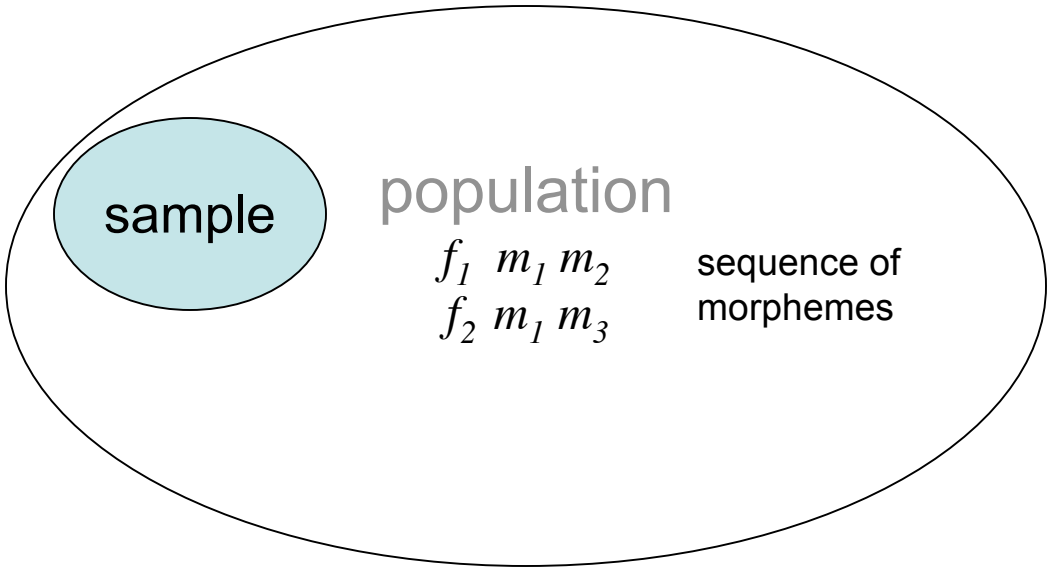
Big Picture

- want results about a sample to generalize to the population



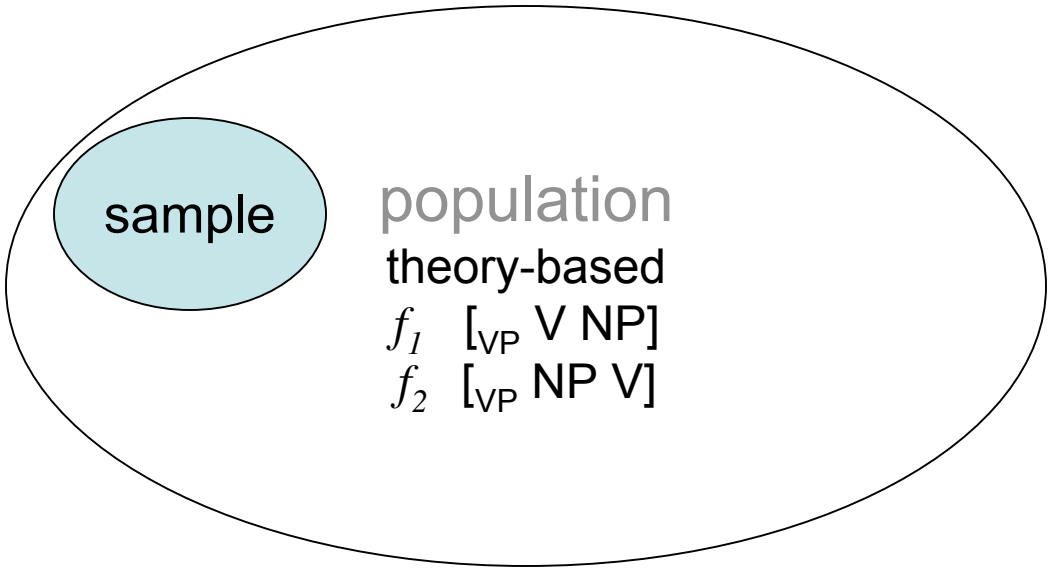
Big Picture

- want results about a sample to generalize to the population



Big Picture

- want results about a sample to generalize to the population

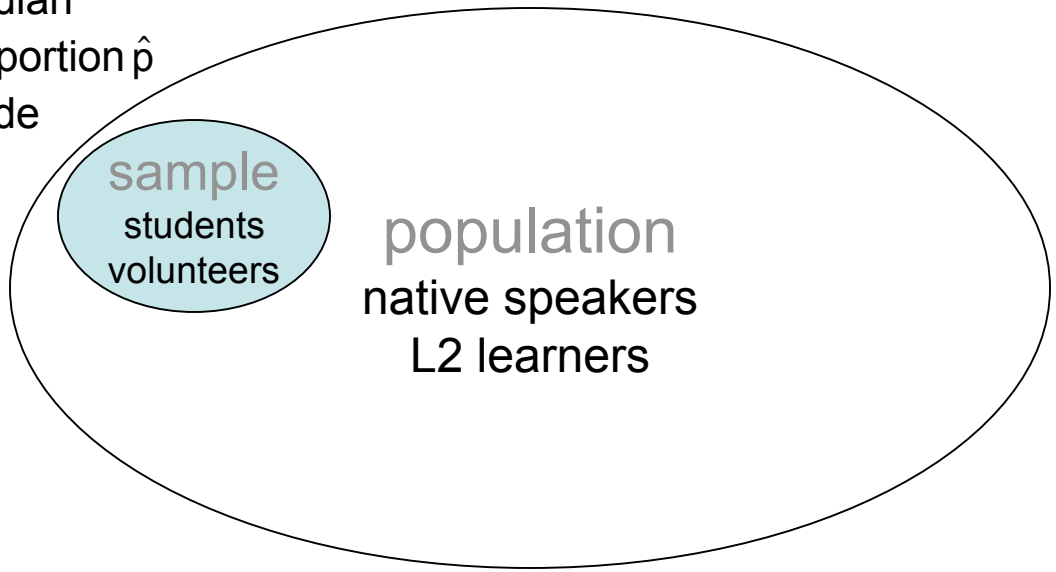


Big Picture

- want results about a sample to generalize to the population
  - **statistical parameters**
    - **sample**
    - variance s^2
    - mean $\bar{x}$
    - median
    - proportion $\hat{p}$
    - mode

population
variance σ^2
mean μ

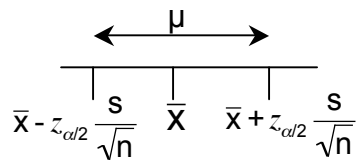
correlation: between more than one random variable, $f(s_1,x_1,s_2,x_2)$
can use it to compute regression
estimate curve parameters: e.g. line: slope and y-intercept



correlation
$\not\Rightarrow$
causality

Big Picture

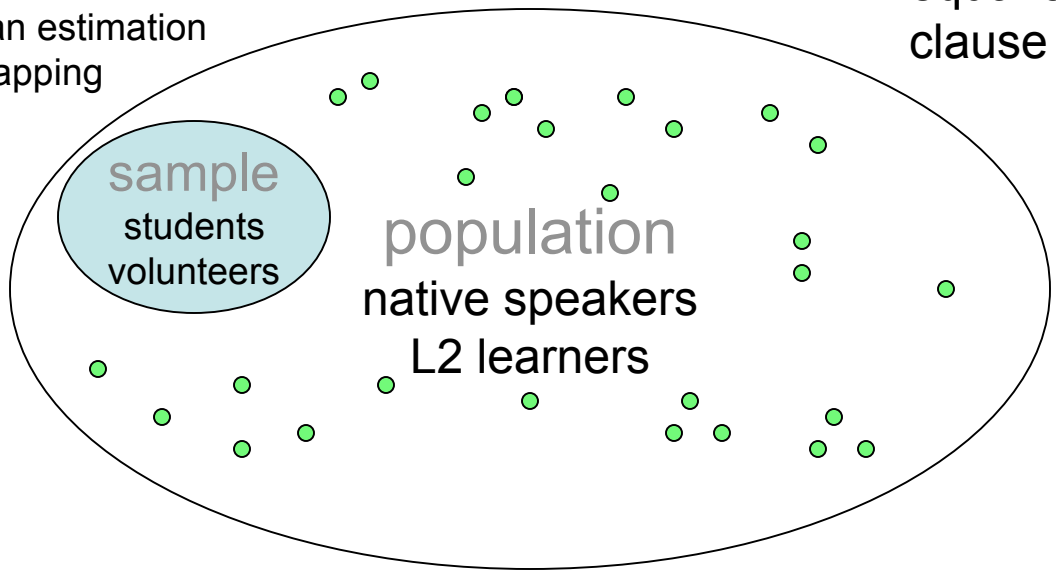
- want results about a sample to generalize to the population
  - **good sample**
  - **random (i.i.d.) – avoid bias**
  - **accurate – large sample size**



hypothesis testing
goodness-of-fit

also do
- Bayesian estimation
- Bootstrapping

equal opportunity clause



Question

- What are the successes of statistical modeling for language?

- *Is language too complicated to be modeled statistically*?

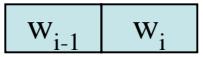
- *Maybe too complicated to be modeled with current symbolic tools as well*?

Refuting Chomsky

- **examples**
  - (1) **colorless green ideas** sleep furiously
  - (2) furiously sleep ideas green colorless
- **Chomsky** (1957):
  - . . . It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, **in any statistical model for grammaticalness**, these sentences will be ruled out on identical grounds as equally `remote' from English. Yet (1), though nonsensical, is grammatical, while (2) is not.
- **idea**
  - (1) is syntactically valid
  - (2) is word salad

Refuting Chomsky

- **examples**
  - (1) colorless green ideas sleep furiously
  - (2) furiously sleep ideas green colorless
- Statistical Experiment (Pereira 2002)



bigram language model

$$p(w_1 \cdots w_n) = p(w_1) \prod_{i=2}^{n} p(w_i|w_{i-1}) \quad .$$

Using this estimate for the probability of a string and an aggregate model with $C = 16$ trained on newspaper text using the expectation-maximization (EM) method (Dempster, Laird, & Rubin, 1977), we find that

$$\frac{p(\text{Colorless green ideas sleep furiously})}{p(\text{Furiously sleep ideas green colorless})} \approx 2 \times 10^5$$

200,000 times as likely!

Thus, a suitably constrained statistical model, even a very simple one, can meet Chomsky's particular challenge.

Statistical Modeling = State of the Art

www.languageweaver.com

Statistical MT System [Spinoff from USC/ISI work]

- "Language Weaver's SMTS system is a significant advancement in the state of the art for machine translation… and [we] are confident that Language Weaver has produced the most commercially viable Arabic translation system available today."

- Metrics: performance determined by competition
  - common test and training data

Statistical Modeling = State of the Art

Google dominates in machine translation tests

update Search giant Google's ambitions to make the Web more international has gotten a slight boost from a U.S. government-run test in which its translation software beat out technology from IBM and academia.

Google scored the highest in Arabic-to-English and Chinese-to-English translation tests conducted by the National Institute of Science and Technology. Each test consisted of translating 100 articles from Agence France Presse and the Xinhua News Agency dated from Dec. 1, 2004, to Jan. 24, 2005. The results were posted earlier this month.

Statistical Modeling = State of the Art

- *Is the bar set too low?*

update Search giant Google's ambitions to make the Web more international has gotten a slight boost from a U.S. government-run test in which its translation software beat out technology from IBM and academia.

Google likely benefited from its huge store of source material. Generally speaking, translation software improves as more data gets fed to it. Through its search operations, Google has amassed billions of translated Web pages.

Google's machine translation wasn't perfect, but it was well ahead of the competition. On a scale from zero to one, the company's software scored 0.5137 on the Arabic tests and 0.3531 on the Chinese tests. The University of Southern California's Information Sciences Institute came in second with a 0.4657 on Arabic tests and 0.3073 on Chinese. IBM scored 0.4646 on Arabic and 0.2571 on Chinese.

Statistical Modeling = State of the Art

- Example:

"Nobody in my team is able to read Chinese characters," says Franz Och, who heads Google's machine-translation (MT) effort. Yet, they are producing ever more accurate translations into and out of Chinese - and several other languages as well.