Theories and statistics

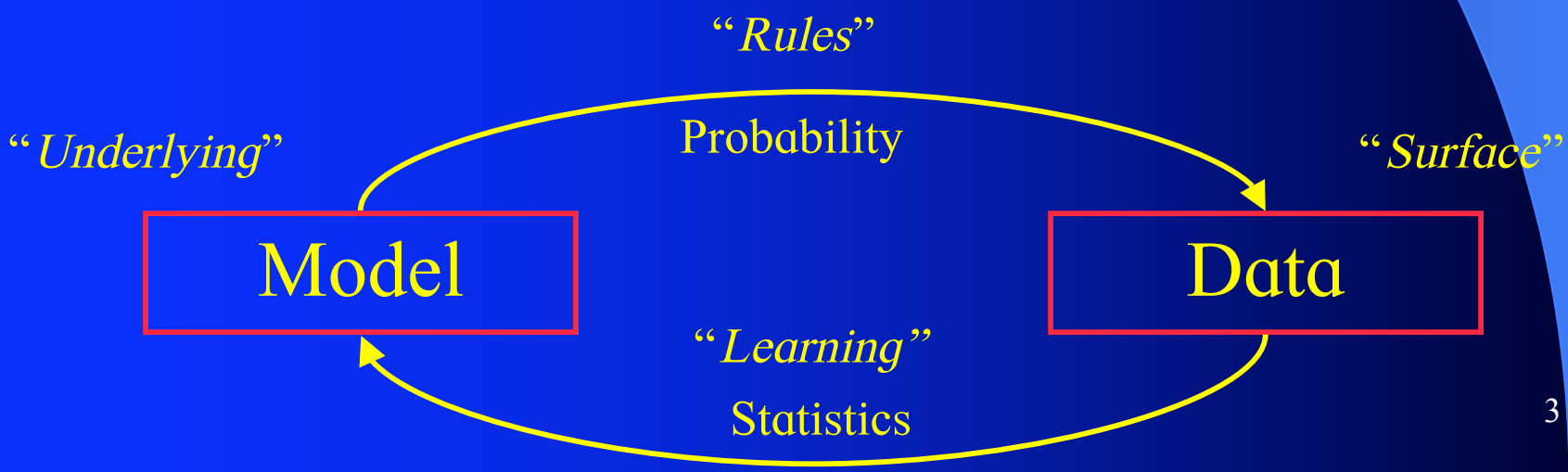
Ying Lin

Statistics

- Why do statistics: want to draw some conclusion from data analysis
  - Fundamentally, not all that different from what linguists want to do
- The main difference between statistical inference and the way linguists draw conclusions
  - Reasoning based on numbers
  - Uses the language of probability

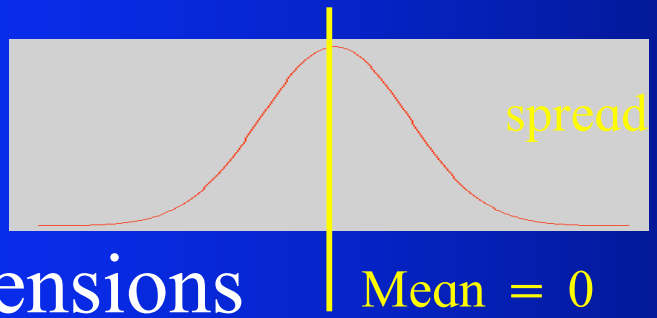
Probability and statistics

- Probability is essentially a coherent way of manipulating numbers
  - Model induces a distribution over data
  - This type of view is actually called "generative model" in statistics -- needs theory!



Example: a small building block in statistical inference

- Normal distributions: rough model of variability in measurements / errors



- Some extensions
  - Adding up the same kind of squared errors -- Chi-square distribution (Mike's talk)
  - Letting the spread of normal be Chi-square -- "Students' t"-distribution

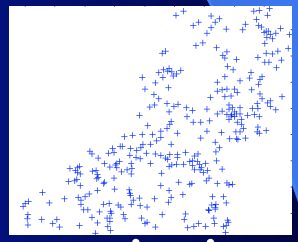
General domain theories: additive models

- Look at ANOVA again:
  Response - grand mean = effect + error

  Total variance — Variance due to effect — Variance due to error
  - If there is something going on, then less variance is due to the error than the effect
  - Inference: divide the two variances (both Chi-square) --> a F-distribution
- Set up the right additive models for specific kinds of data: issue of sampling (Natasha)

Need for domain-specific theories: Dutch example

- Background:
  - Dutch speakers are aware of non-standard varieties that use different word orders
  - Data gathered from carefully selected speakers from different regional varieties



- Goal:
  - Identifying major trends of syntactic variation
  - Testing theories of syntax and their predictions about word order

What to do with the Dutch data set?

(omdat Jan moet hebben gezwommen: “because John must have swim”)

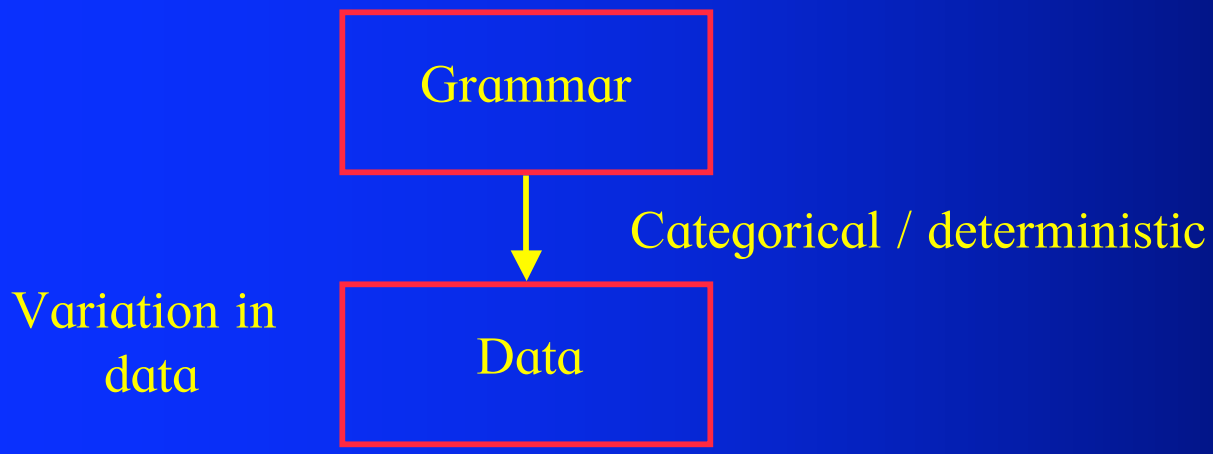
Must have particle-swim

Test sentence	S1	S2	S3	S4	S5
moet hebben gezwommen	*	*	?	?	OK
moet gezwommen hebben	*	*	?	?	OK
gezwommen moet hebben	*	?	OK	OK	OK
gezwommen hebben moet	OK	OK	OK	*	*
hebben gezwommen moet	*	*	*	?	*
hebben moet gezwommen	*	*	*	?	*

Magnitude estimates also available

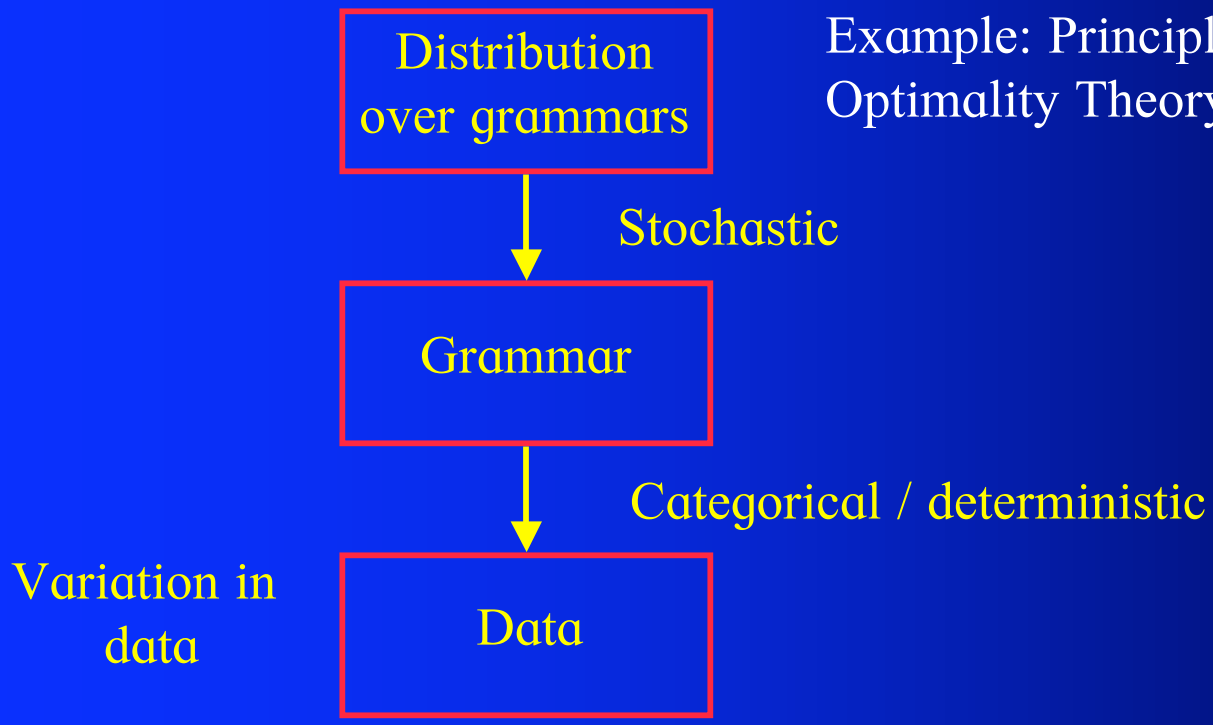
Towards theoretically informed analysis

- Parametric variation from a hierarchical generative model



Towards theoretically informed analysis

- Parametric variation from a hierarchical generative model



Example: Principle & Parameters, Optimality Theory

Message

- Theorist can probably make use of a wider range of data when equipped with tools.
- The language of probability is rich and can be combined with theories in a meaningful way.
- Data analysis needs to be informed by theories for drawing complex conclusions.
  - Development of tools needs help from theories too!