

# Data Generating Processes and Statistical Modeling

Jeffrey B. Arnold

January 28, 2016

# Ontological Interpretations of Probability

DID THE SUN JUST EXPLODE?  
(IT'S NIGHT, SO WE'RE NOT SURE.)

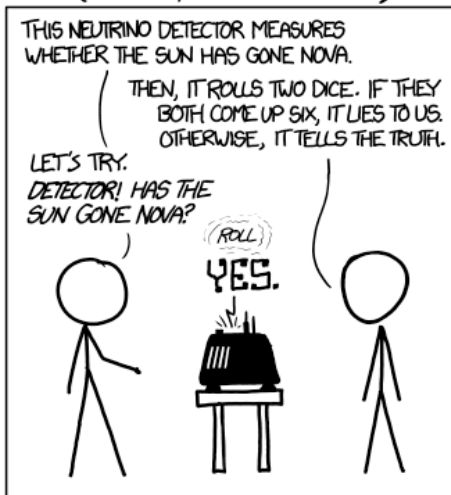
THIS NEUTRINO DETECTOR MEASURES  
WHETHER THE SUN HAS GONE NOVA.

THEN, IT ROLLS TWO DICE. IF THEY  
BOTH COME UP SIX, IT LIES TO US.  
OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY.

DETECTOR! HAS THE  
SUN GONE NOVA?

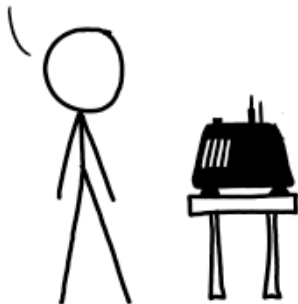
ROLL  
YES.



# Ontological Interpretations of Probability

FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT  
HAPPENING BY CHANCE IS  $\frac{1}{36} = 0.027$ .  
SINCE  $p < 0.05$ , I CONCLUDE  
THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU \$50  
IT HASN'T.



# Ontological Interpretations of Probability

- ▶ Types of Probability
  - ▶ Frequentist (Classical)
  - ▶ Bayesian (Subjective)
- ▶ Probability of one-off events
- ▶ Beliefs of actors
- ▶ Mathematically equivalent rules

# Data Generating Processes (DGP)



*[A] data generating process (DGP) is rule or set of rules governing the social or political events that an analyst wishes to study and the rules by which observations of its results come to be represented in a dataset. A DGP ... governs how a factors in a political process are related to each other. SMISS (p. 70)*

# DGP: Stochastic vs. Deterministic

- ▶ **Deterministic:** sufficient and necessary conditions to observe data
- ▶ **Stochastic:** probability of observing the data
- ▶ Why treat DGP as stochastic?
  - ▶ Sampling uncertainty
  - ▶ Theoretical uncertainty
  - ▶ Fundamental uncertainty

# DGP and Statistical Models

- ▶ Stochastic DGP treats data as random variables
- ▶ Theoretically interested in specific parameters of the random variable: expectation
- ▶ *Model-based sampling*: assume a specific parametric probability distribution
- ▶ Problems with *model-based sampling*: DGP adds additional structure not implied by theory

# Statistical Model with a Normal Distribution

$$Y = \mu + \epsilon$$

$$\epsilon \sim N(0, \sigma)$$

$$\mu = \alpha + \beta X$$

- ▶ Setting up a model is “statistical modeling” and defining our DGP
- ▶ How we estimate the parameters in the model  $\alpha, \beta, \sigma$  to make inferences about the DGP is a separate problem.
- ▶  $\epsilon$  is the *stochastic component*
- ▶  $\mu$  is the *systematic component*