GG 501 SPATIAL KNOWLEDGE MOBILIZATION

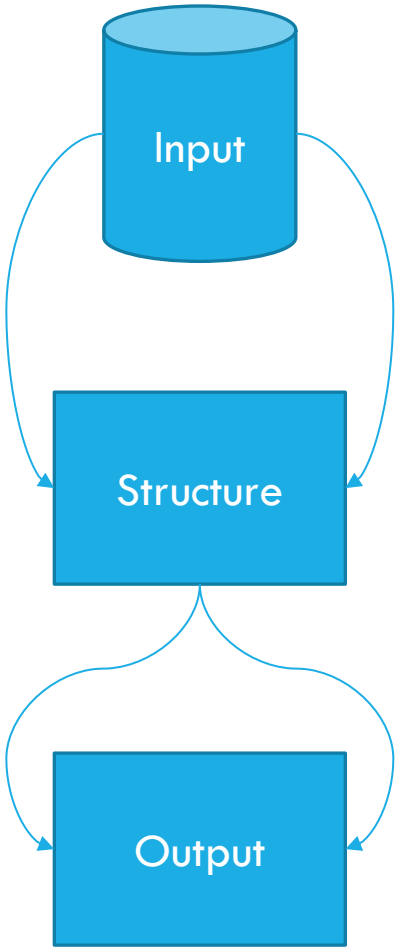
Feb 8: Model data

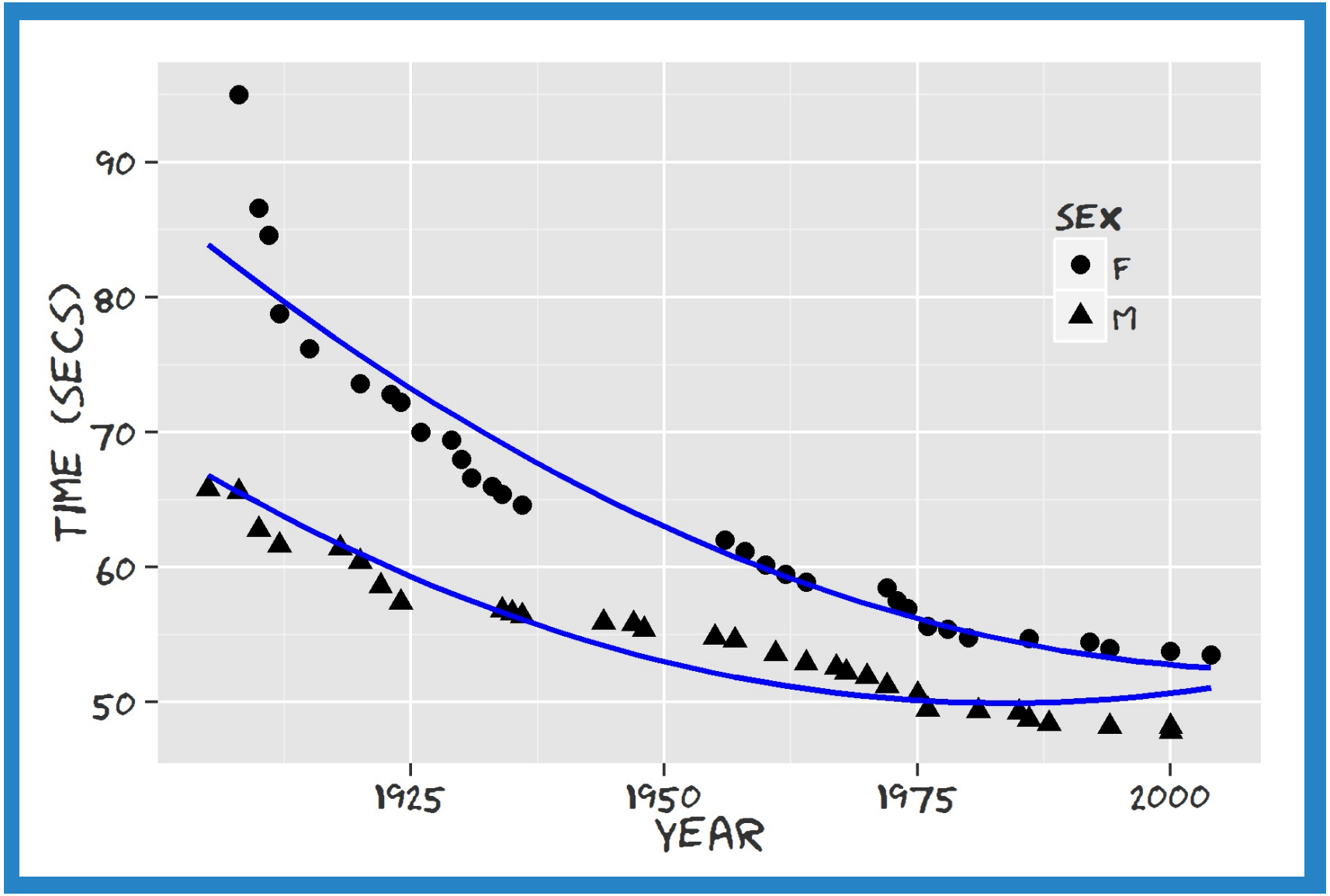
MODELS IN R –

- What do we mean by the word ‘model’?

MODELS COME IN MANY VARIETIES

- Each take some input data
- Attempt to generalize about the underling data-generating-process
- Can be used for a variety of purposes –
  - description
  - explanation
  - prediction






```
time ~ 1 + year + sex + year:sex + I(year^2)
```

STATISTICAL VS PROBABILITY MODEL

- Statistical model
  - Describes one or more variables & their relationship

- Probability model
  - Describes outcome of random **event**
  - Sometimes called a random variable

DESCRIBING RANDOMNESS

- Random event/variable
  - Sample space - what are the possible outcomes?
- Probability model
  - Assign probability to each member of sample space
    - For a coin toss this is 0.5 for heads & 0.5 for tails
- Purely random
  - Probability model contains all the information
  - **No explanatory variables needed to account for variation**

SETTINGS FOR PROBABILITY MODELS

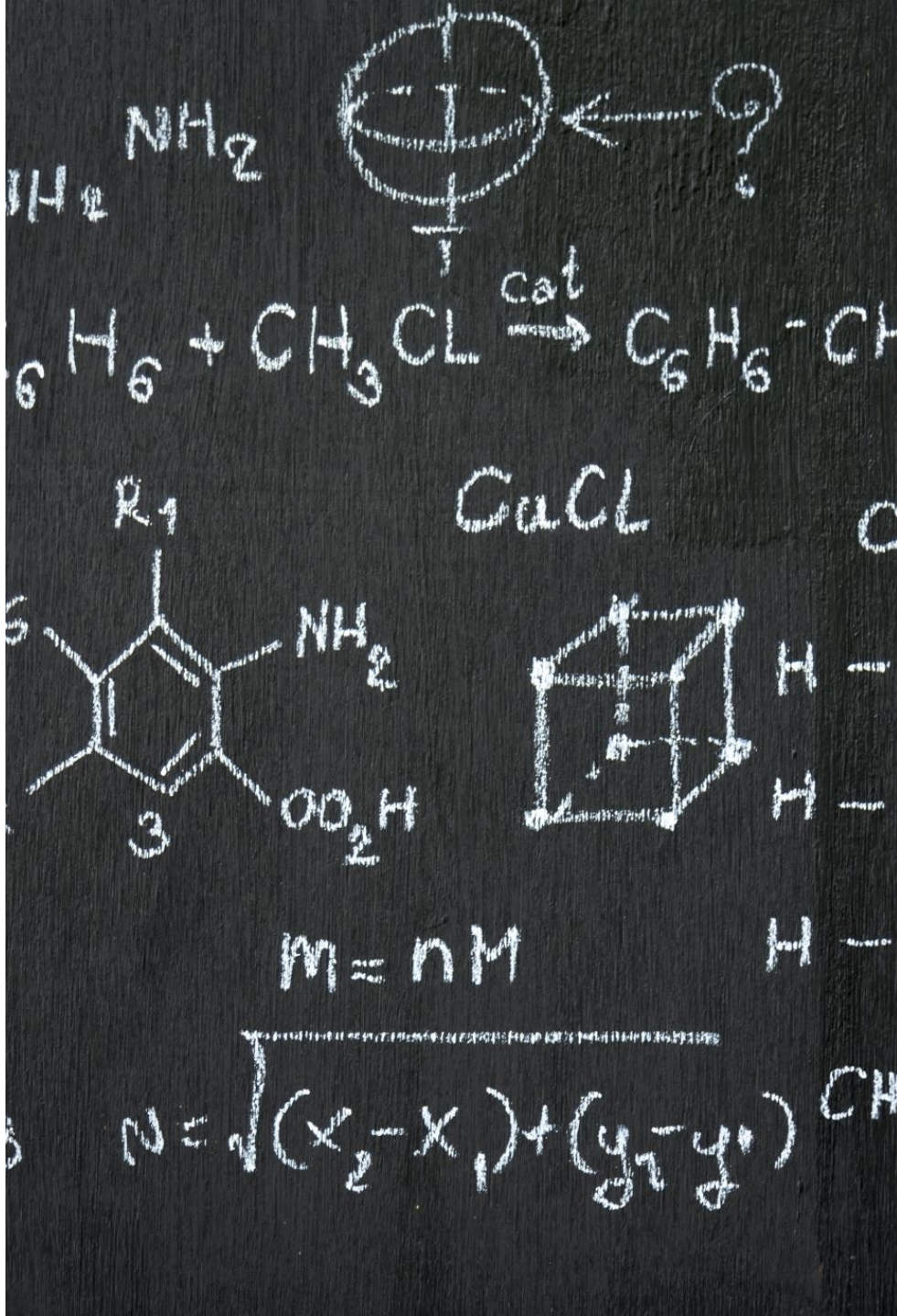
- How is the event configured/measured/represented?
- Examples
  - Number of radioactive particles per minute
  - Student test score on standardized test
  - Number of blood vessels in microscope slide
  - Number of people who support a candidate in a random sample
- Must pick form of model that fits setting
  - Combines expert knowledge & probability calculus

DISCRETE VS CONTINUOUS

- Helpful to distinguish between two kinds of sample spaces
  - Discrete numbers - outcome of rolling a die
  - Continuous numbers - any value in a range
- Possible to assign a probability to each outcome for discrete numbers
- Possible to assign probability to **range** of outcomes for continuous numbers

PROBABILITY DENSITY

- Can assign **probability density** to each outcome by dividing probability by extent of range
  - Usually treat this probability density as function of value of random value: *p(x)*
- Often use probabilities & probability densities in a similar way
  - For discrete sample space, assigned probabilities over all the members of space must add to 1
  - For continuous sample space, *integral* over assigned probability over possible values must be 1

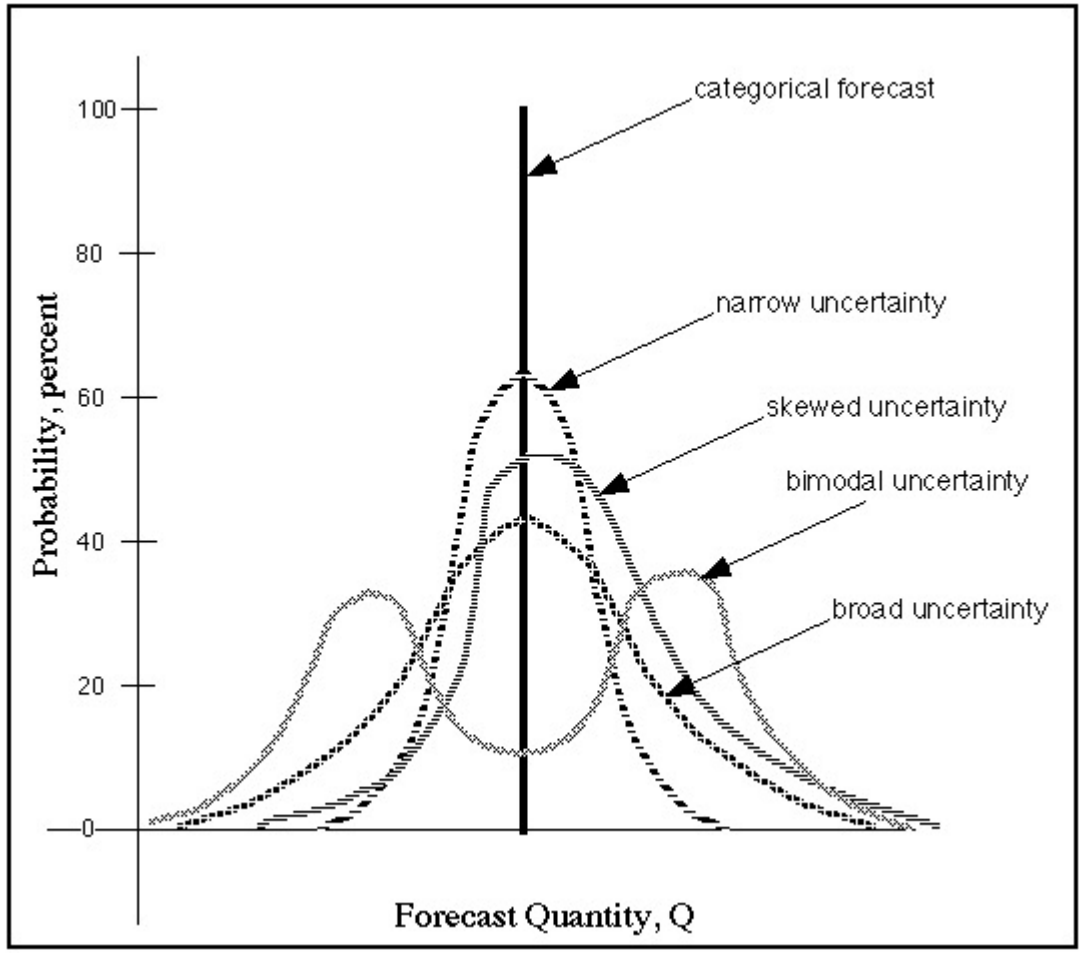


MULTIPLE VIEWS OF PROBABILITY

- Frequentist view of probability
  - Describe how often outcomes occur
    - Example - 100 coin flips should lead to 50 heads
  - Based on large number of possible trials

- Subjectivist view of probability
  - Encodes modeller's assumptions/beliefs
  - Assess degree of belief
  - Probability is assigned to a hypothesis

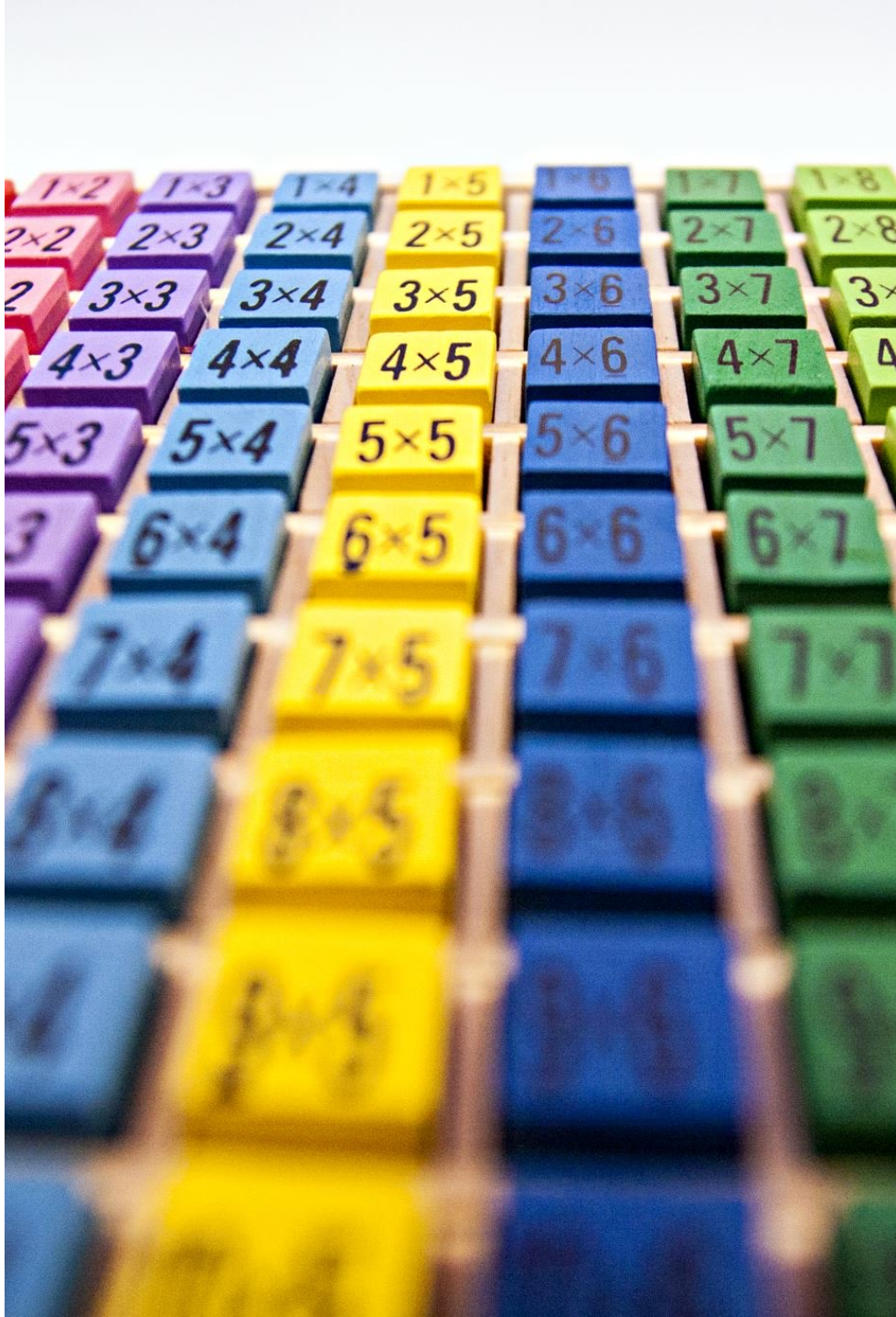
"IT WILL SNOW TODAY…"



http://www.hpc.ncep.noaa.gov/wwd/winter_wx.shtml

STANDARD PROBABILITY MODELS

- Small set of standard probability models apply to wide range of settings
- Don't need to derive them!
- Each model has parameters that need to be adjusted
  - Parameters are similar to coefficients from regression models

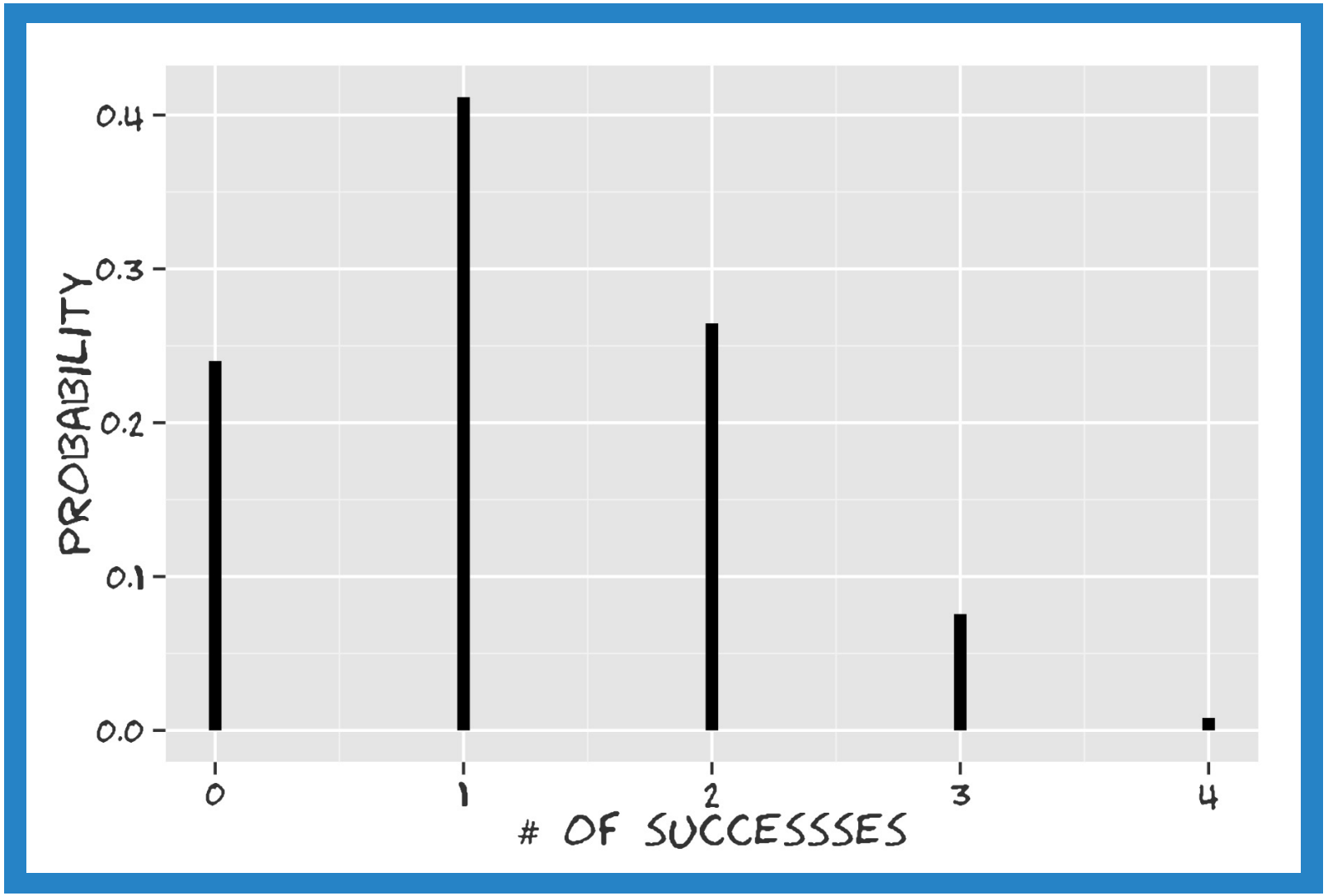


DISCRETE

- Equal probabilities
  - Examples - die toss, coin flip, distributions of ranks of any continuous variable
  - Parameter
    - size - how many possibilities

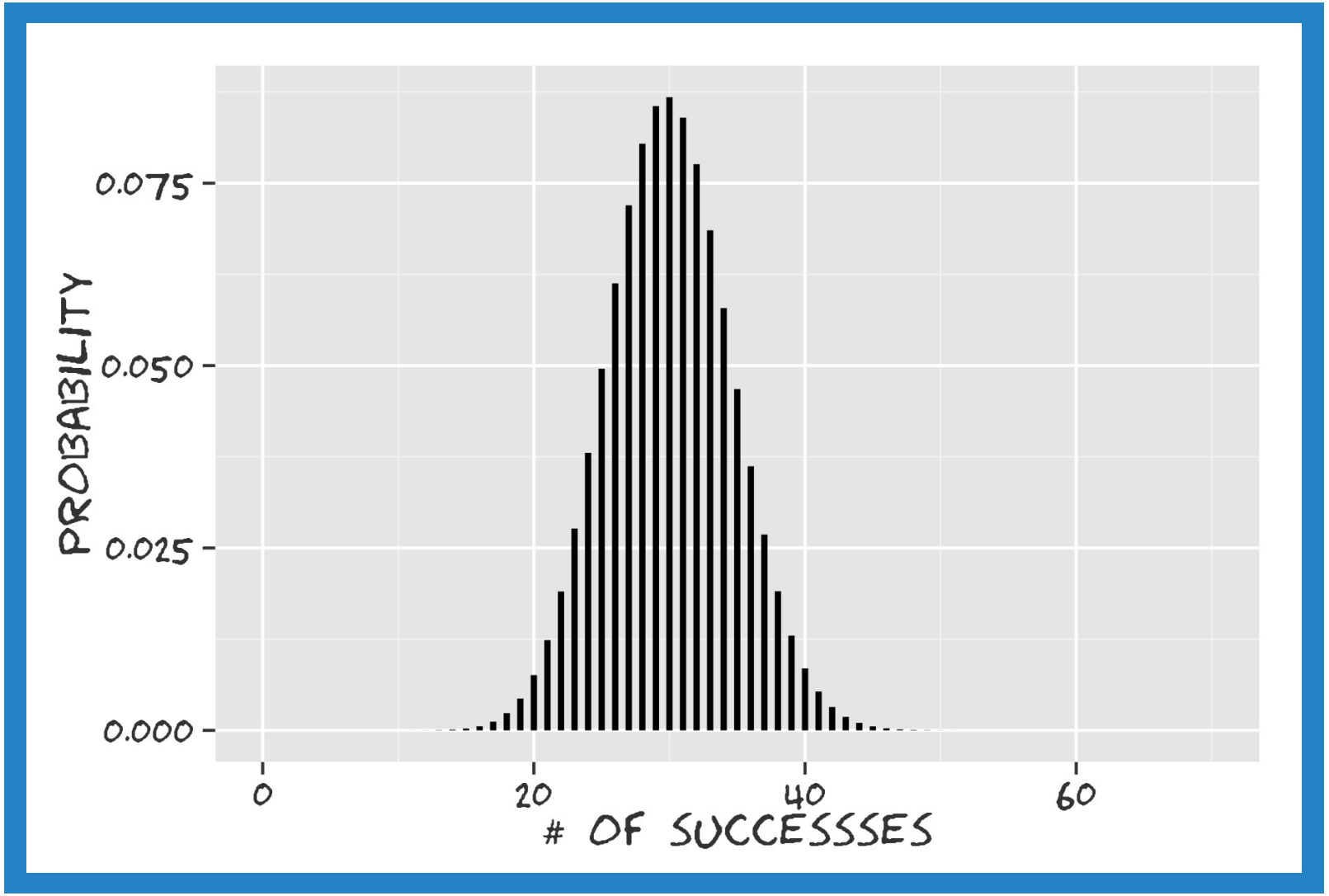
DISCRETE

- Binomial
  - Example - trials of coin flip where outcome is count of “successes” or “heads” or “1s
  - Parameters
    - size - number of trials
    - prob - probability of success on each trial



Binomial model

n=4, p=0.30

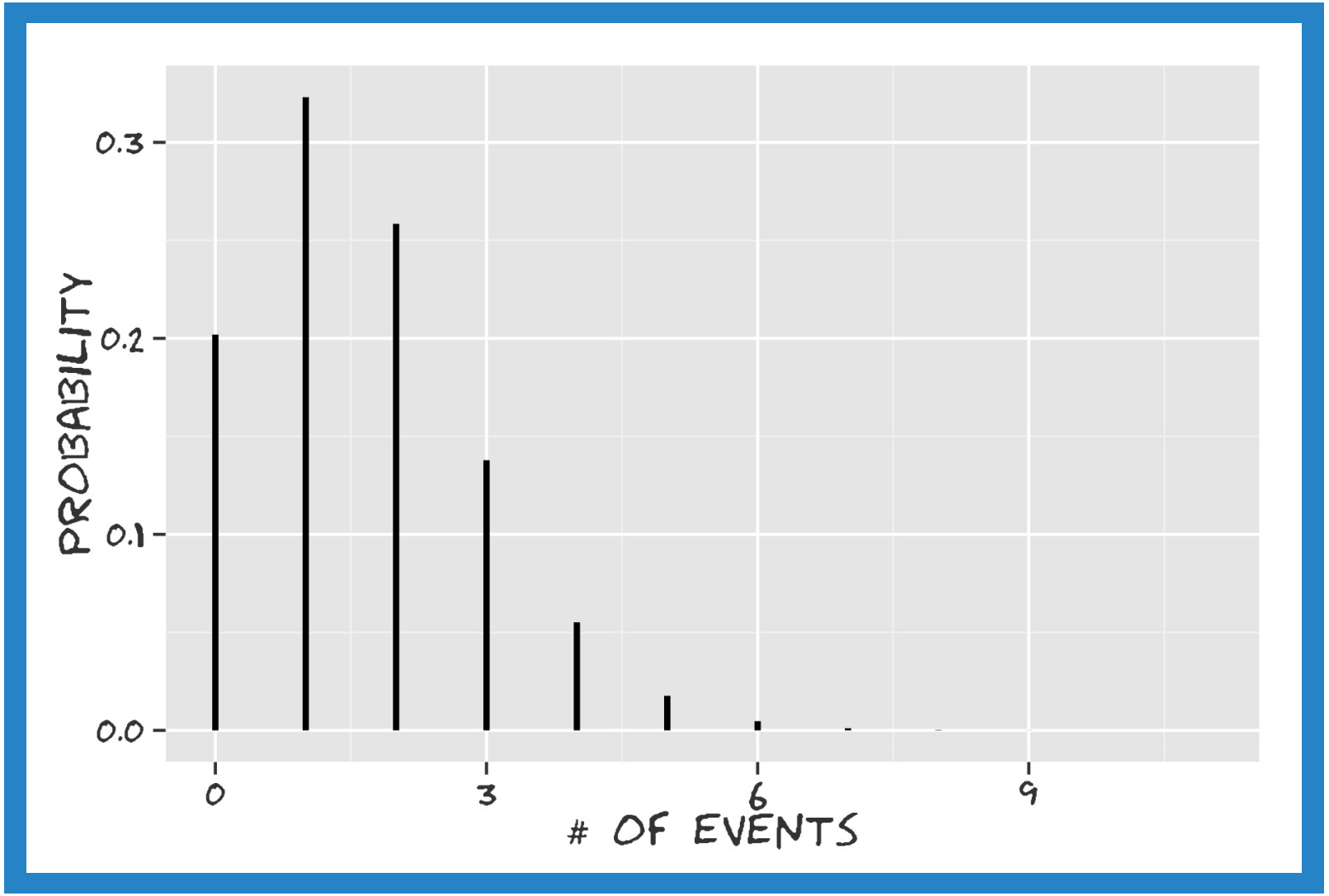


Binomial model

n=100, p=0.30

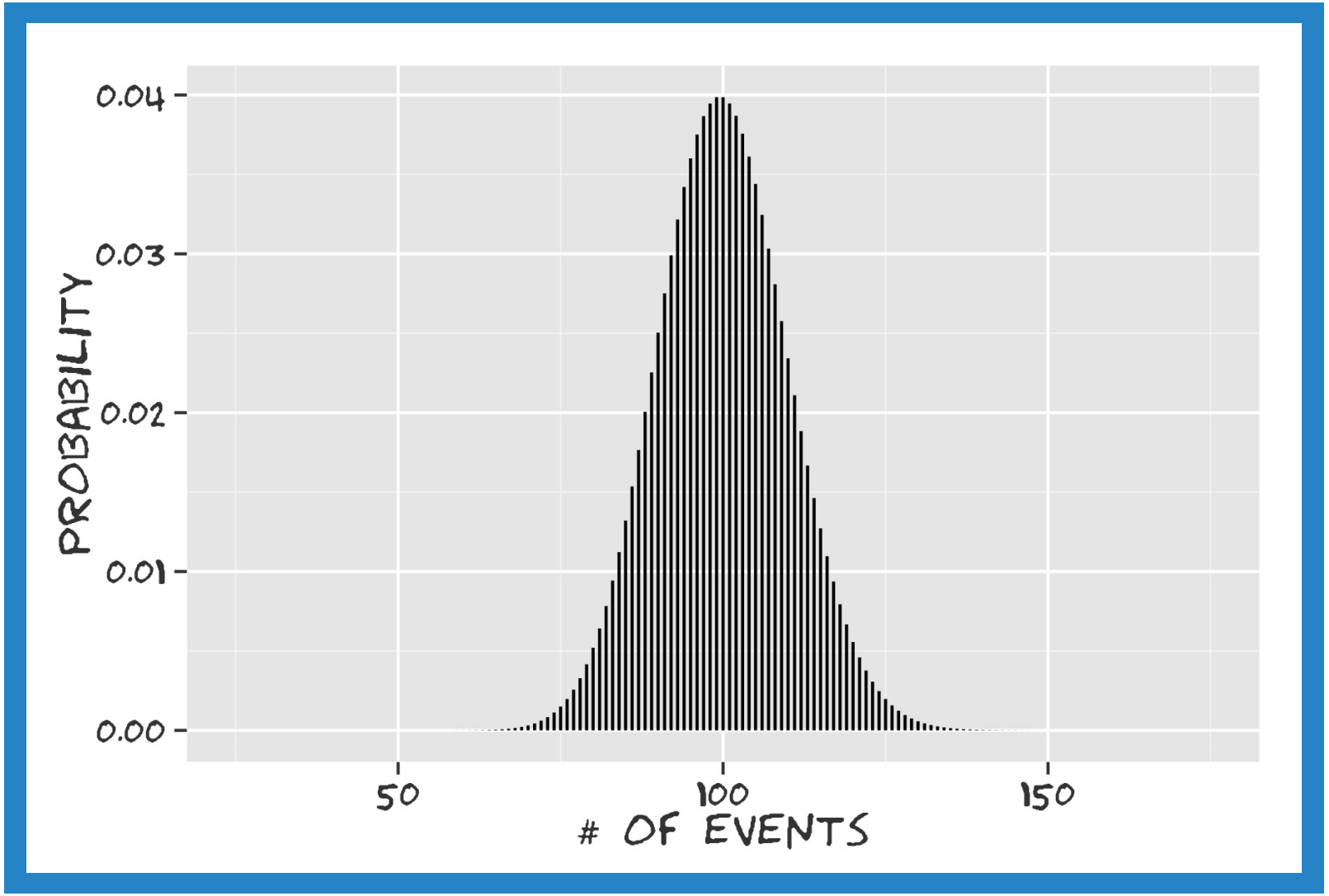
DISCRETE

- Poisson
  - Number of events that happen in a given time
  - Example - number of cars passing by a point, number of shooting stars in minute, number of snowflakes that land on a glove in a minute
  - Parameter
    - lambda - mean number of events



Poisson model

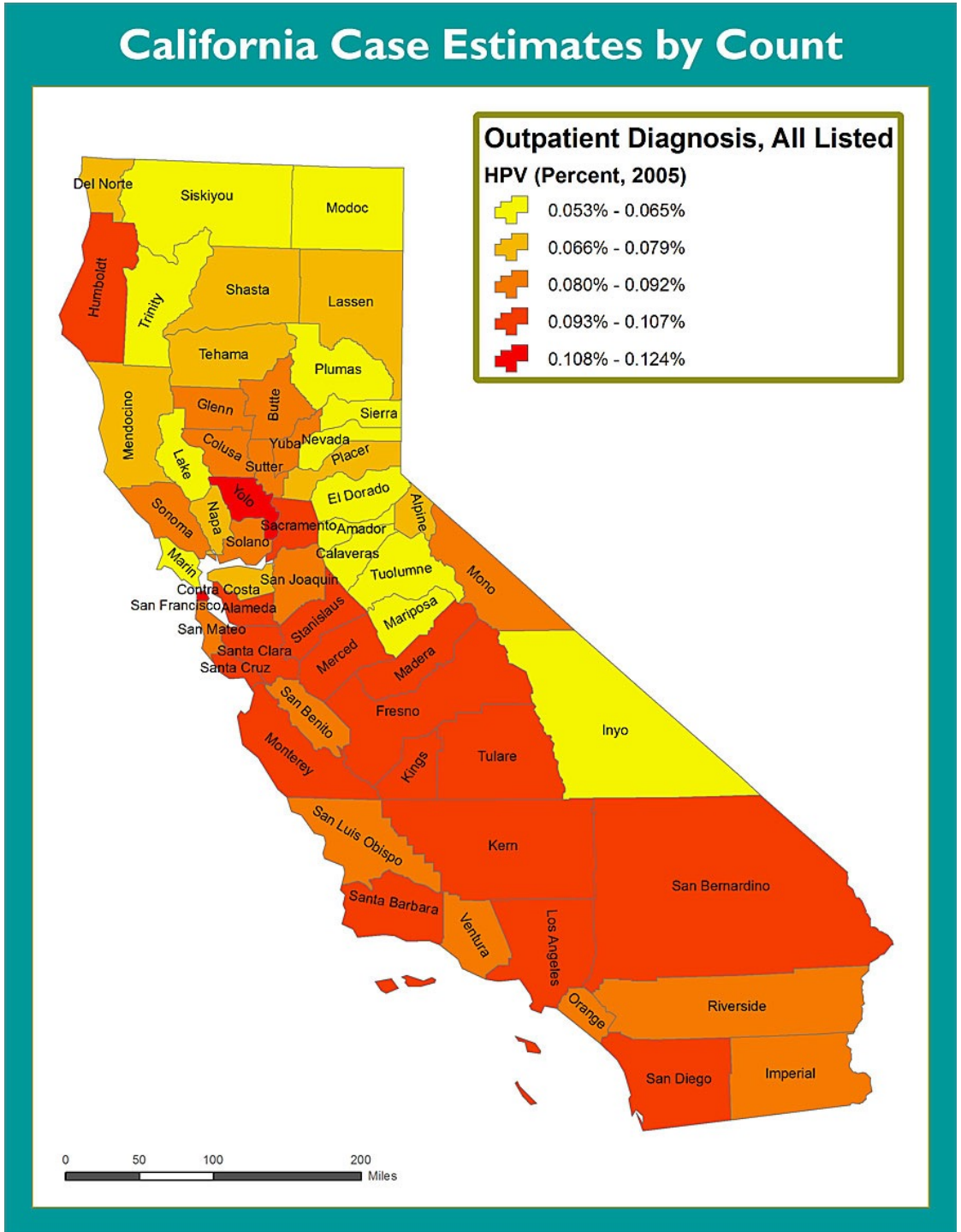
rate (lambda)=1.6



Poisson model

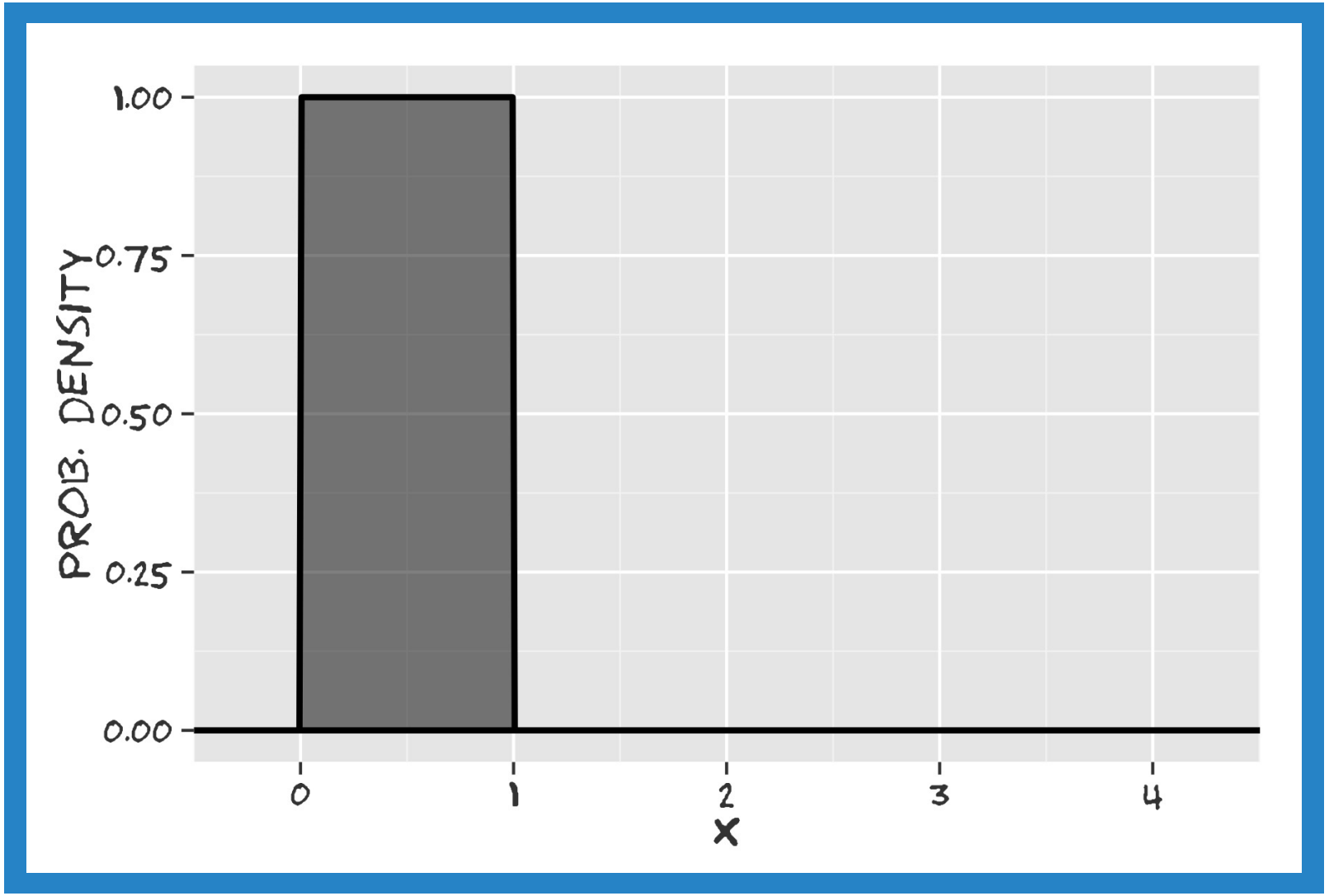
rate (lambda)=100

Map of California human papilloma virus (HPV) cases by outpatient diagnosis in each county



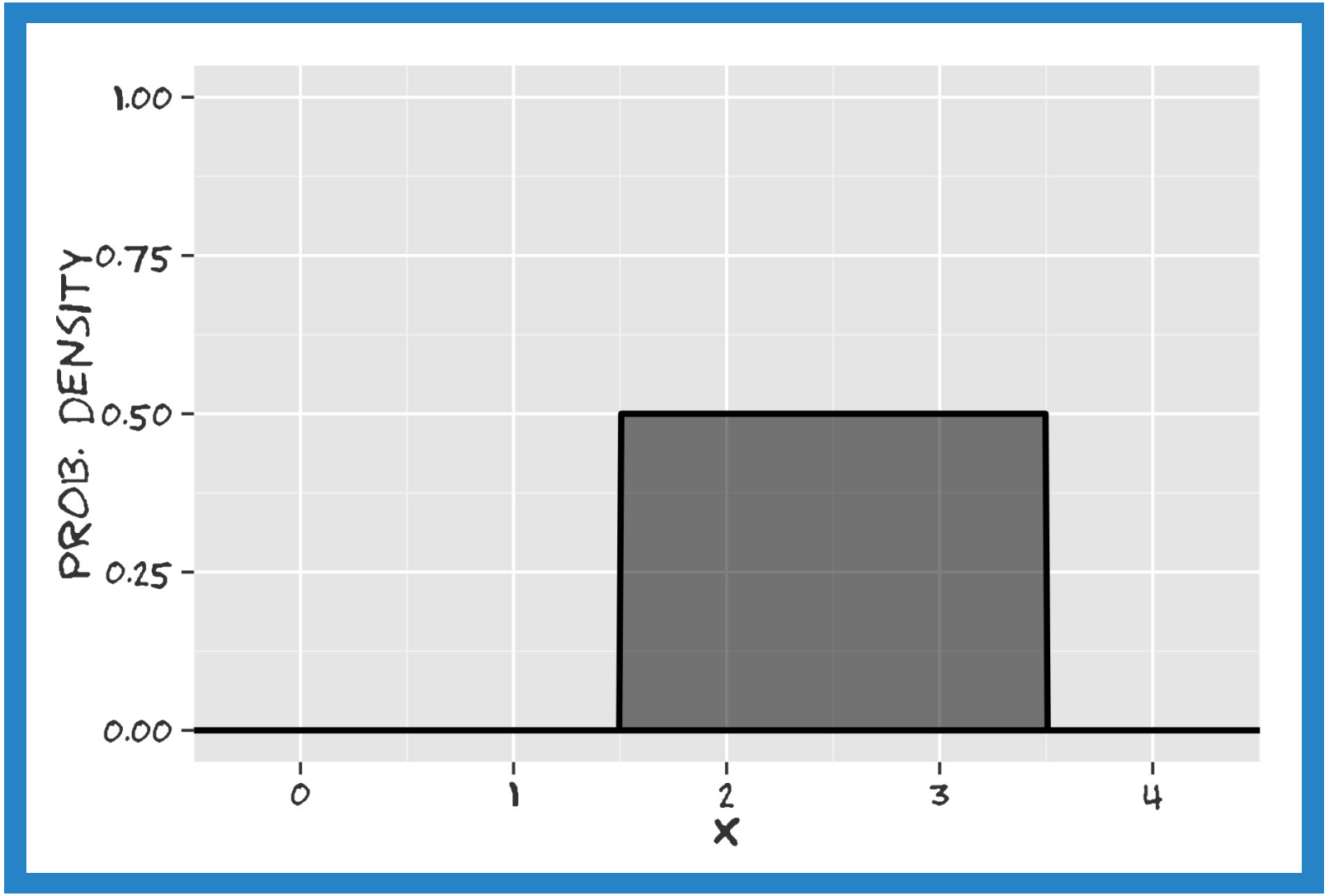
CONTINUOUS

- Uniform
  - Parameters: Max and min
  - Models: spinners (angles in 2-d, but not higher), p-values under the Null Hypothesis
- Normal (gaussian)
  - Parameters: mean and sd
  - Models: your general purpose model



Uniform model

min=0, max=1



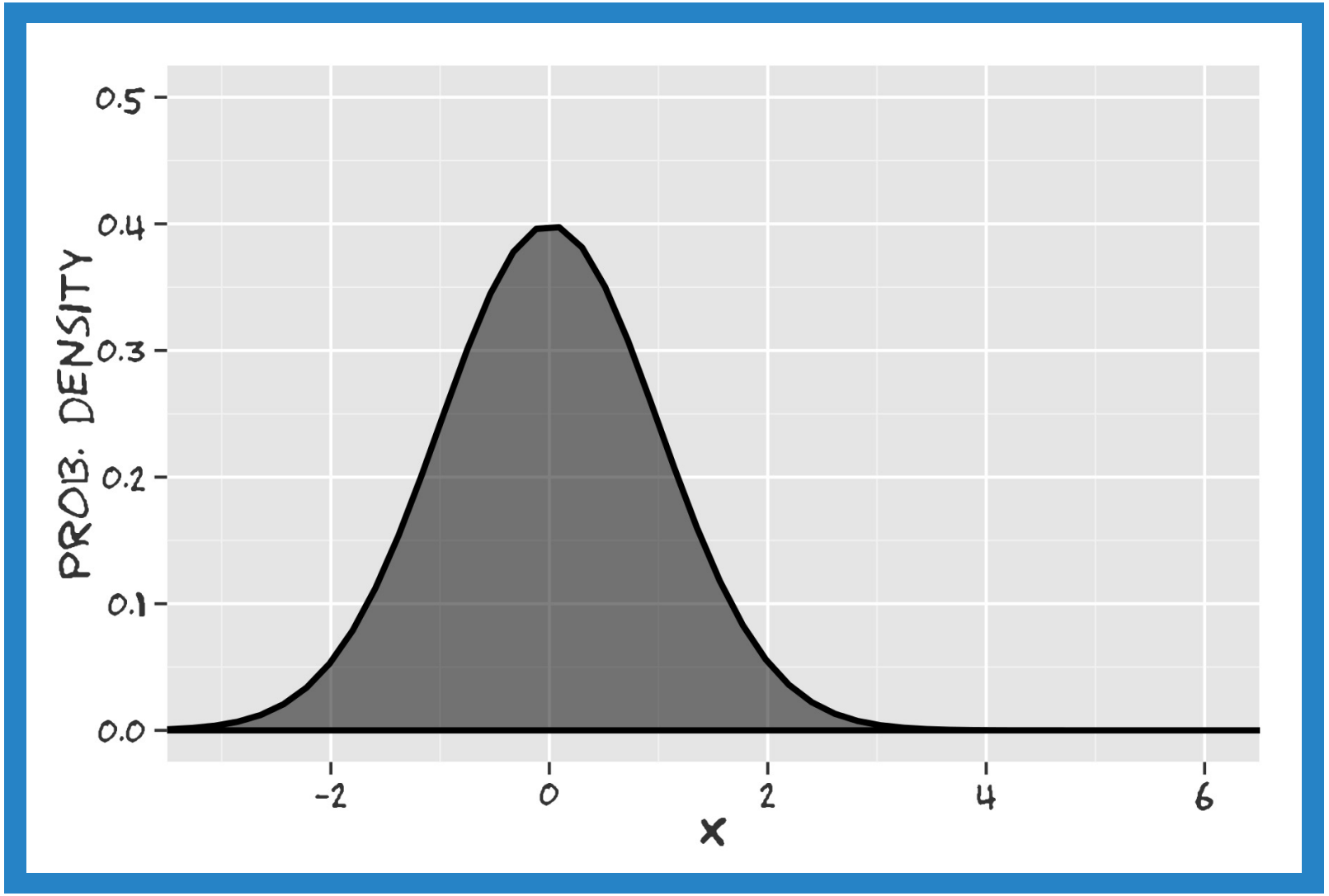
Uniform model

min=1.5, max=3.5

CONTINUOUS

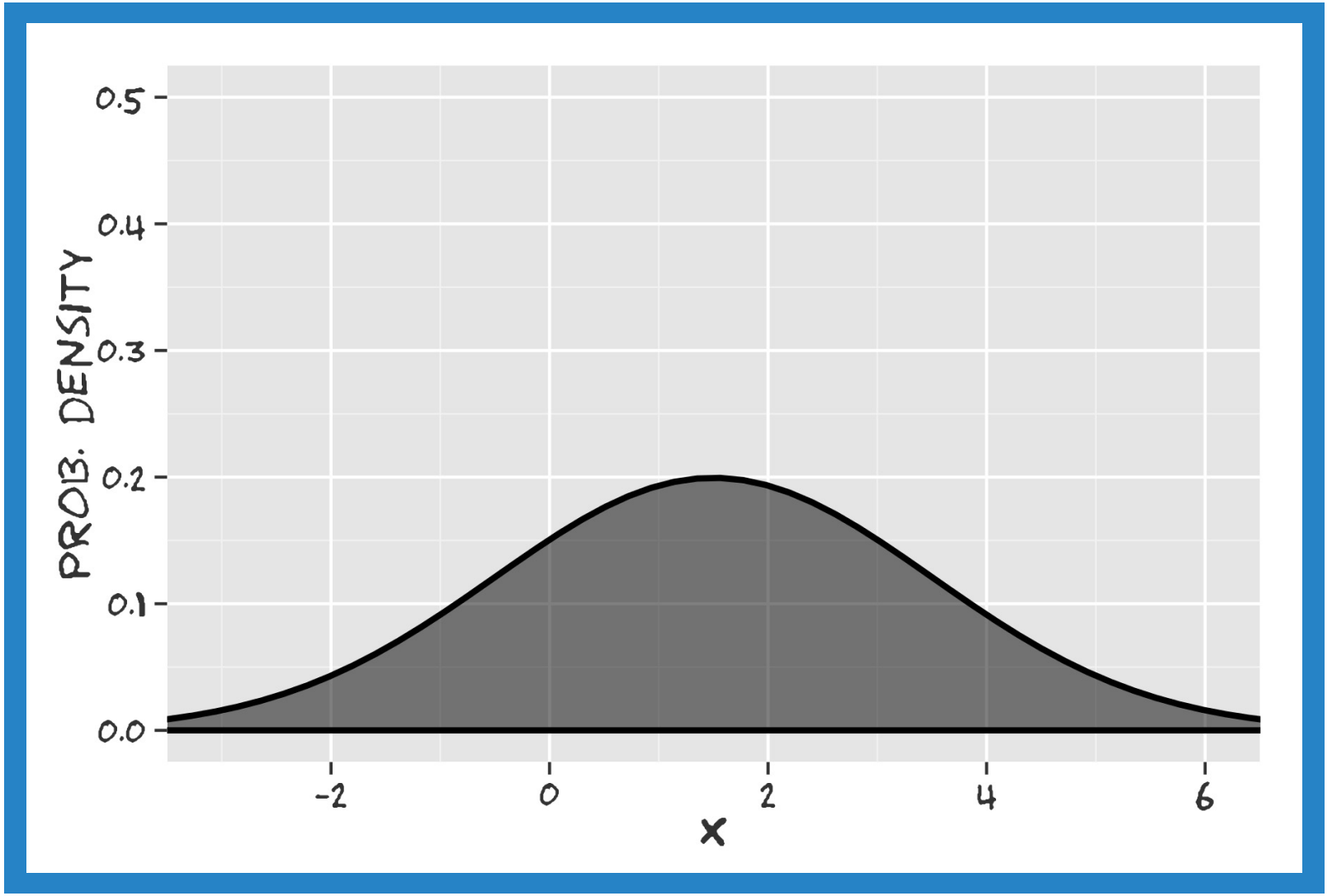
- Uniform
  - Parameters: Max and min
  - Models: spinners (angles in 2-d, but not higher), p-values under the Null Hypothesis

- Normal (gaussian)
  - Parameters: mean and sd
  - Models: your general purpose model



Normal model

mean=0, sd=1



Normal model

mean=1.5, sd=2

WHAT CAN WE DO WITH PROBABILITY MODELS?

Percentiles: what is the range of values within some probability range?

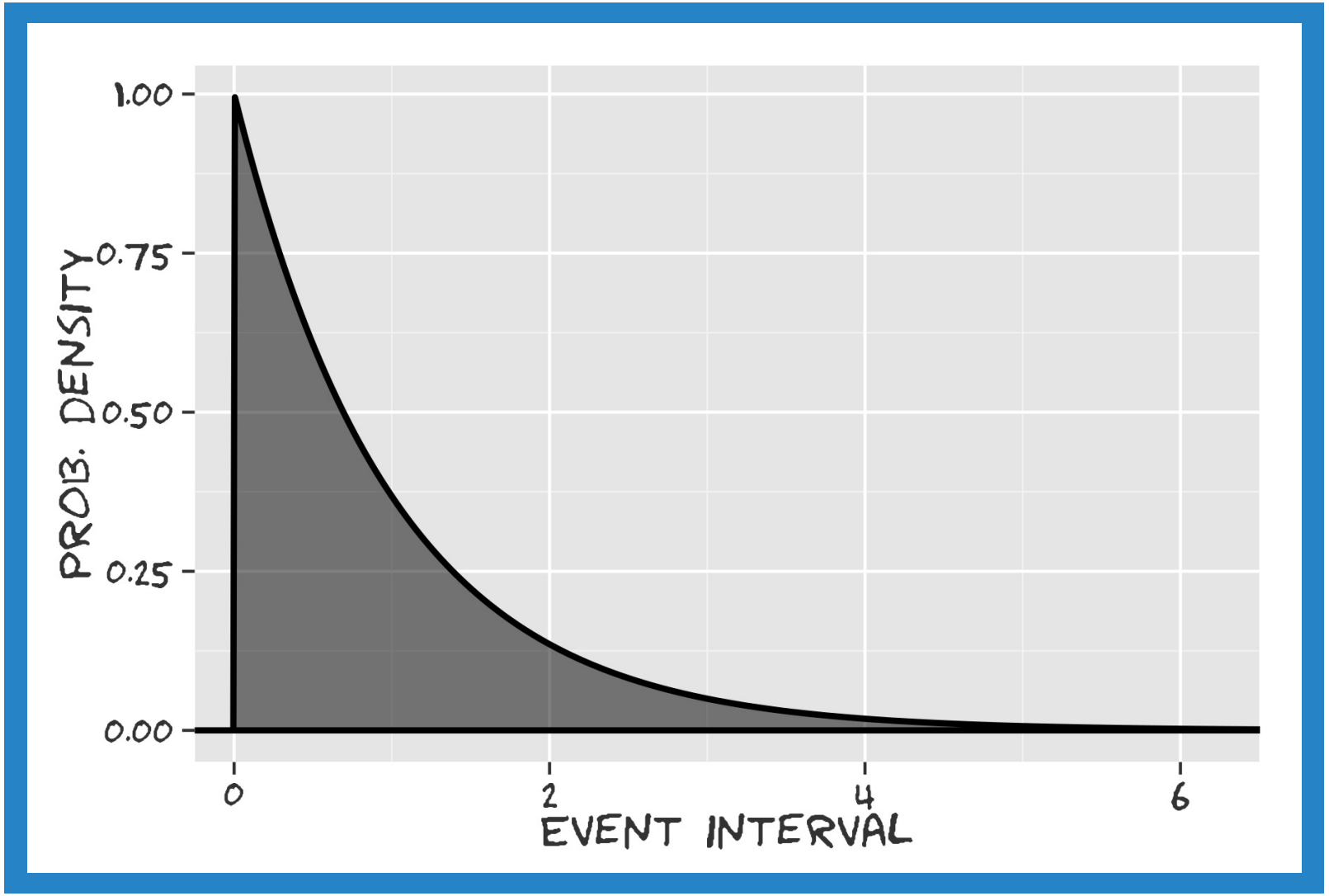
- e.g., 90th percentile: value of the random variable such that 90% of the time values will be equal or smaller

Quantiles: What is the percentile of a given outcome?

- e.g.: how unusual is this observation given the underlying probability model?

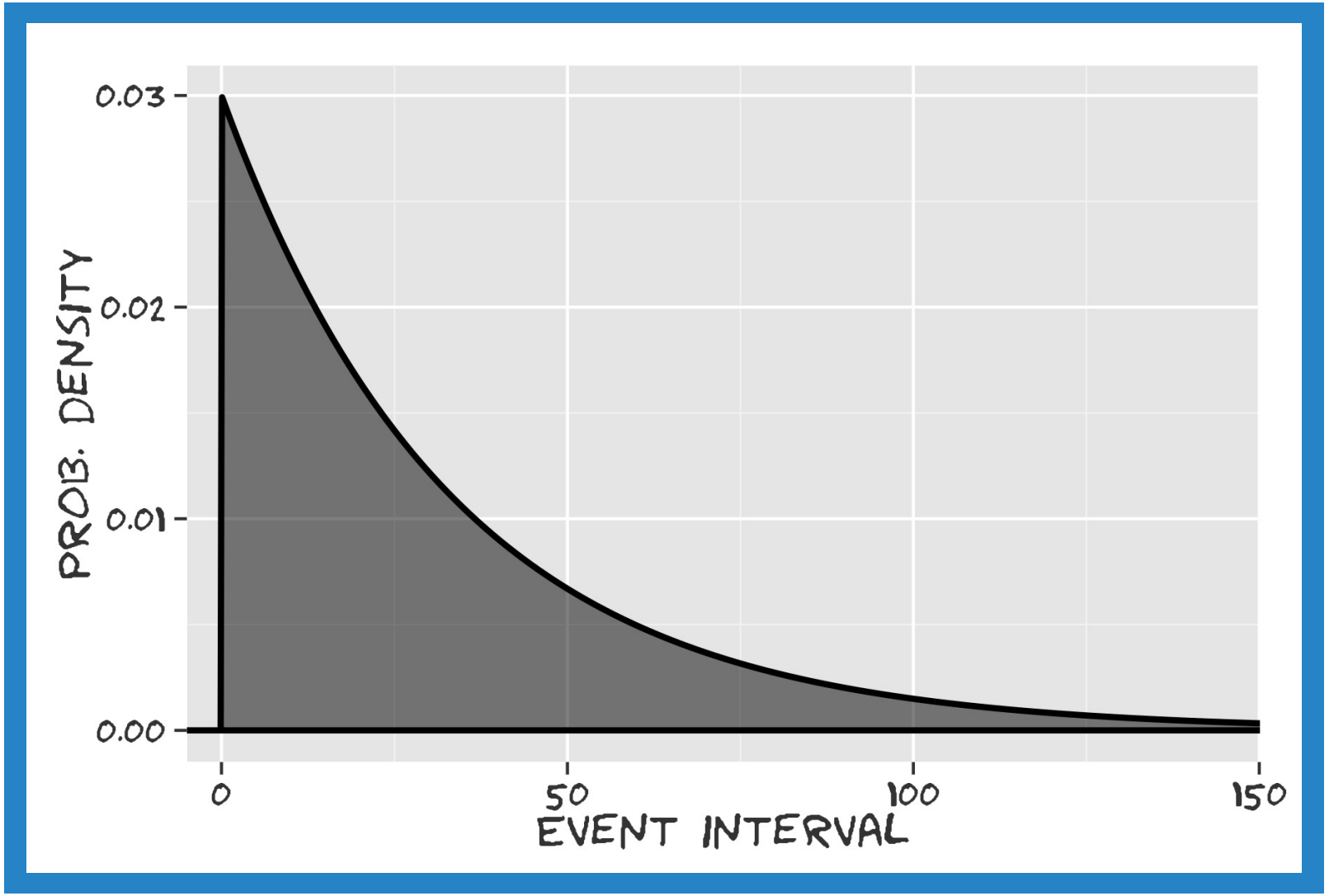
CONTINUOUS

- Exponential
- Times between random events (earthquakes, 100-year storms)
- Parameter: rate (the mean time is 1/rate)



Exponential model

rate=1

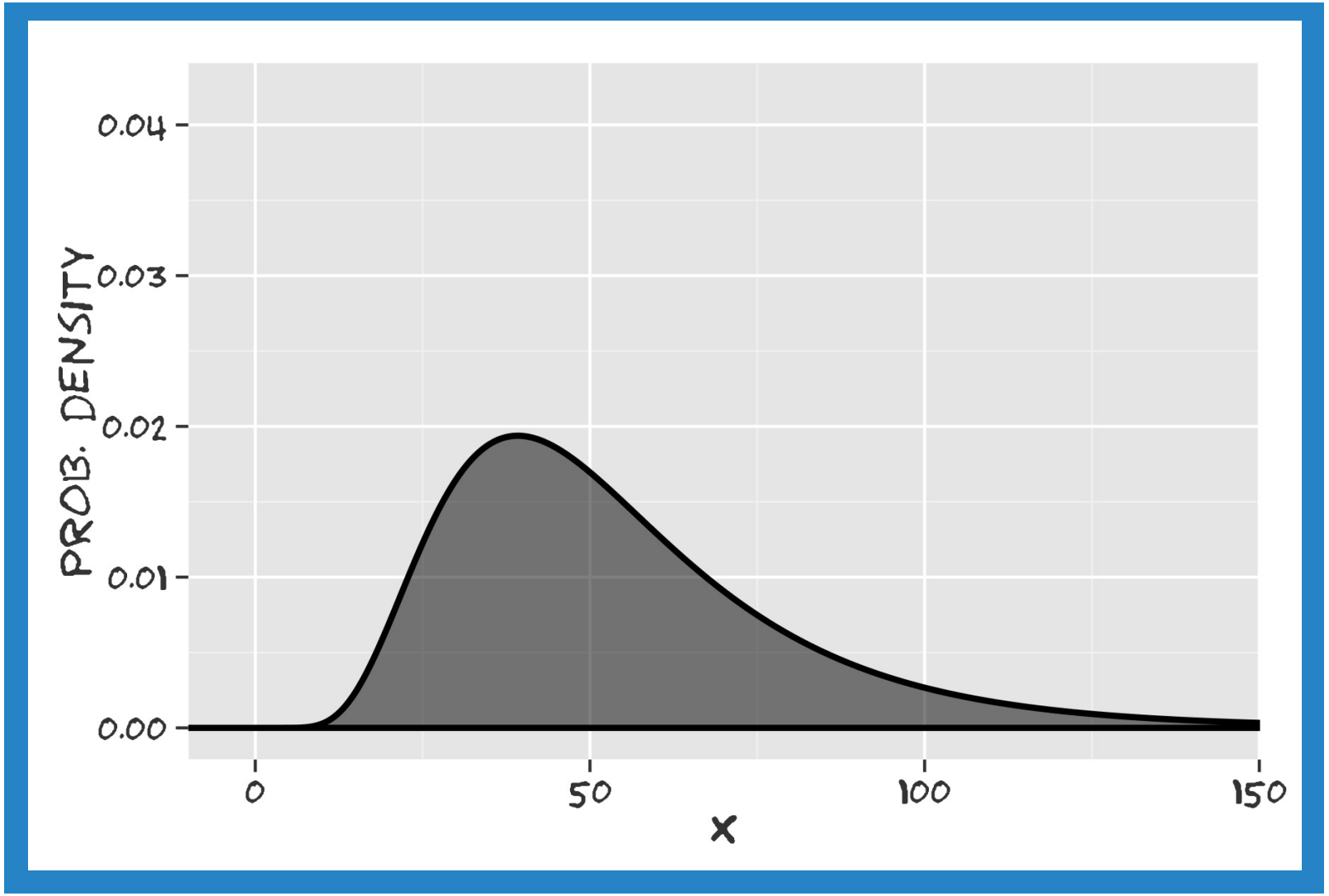


Exponential model

rate=0.03

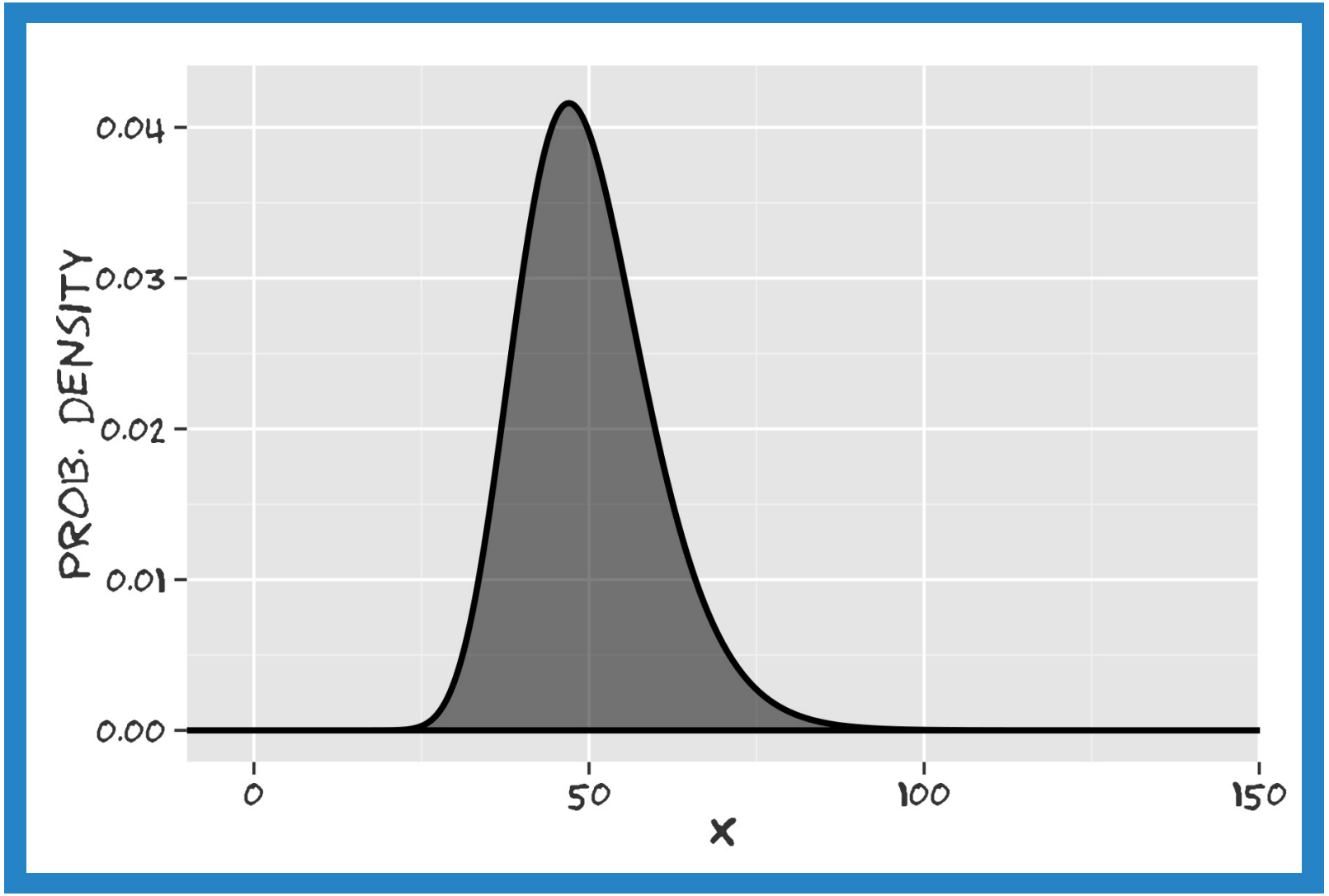
CONTINUOUS

- Lognormal
  - Parameters: mean and sd of the log of the values
  - Models: useful when skew is important



Lognormal model

mean=3.89, sd=0.47



Lognormal model

mean=3.89, sd=0.20

WHY THE NORMAL DISTRIBUTION?

- Many chance phenomena are at least *approximately* described by a normal probability density function
- Example
  - Collect 1000 snowflakes & weigh them, would find distribution of weights accurately described by a normal curve
  - Measure the strength of bones in wildebeests, again likely to find they are normally distributed

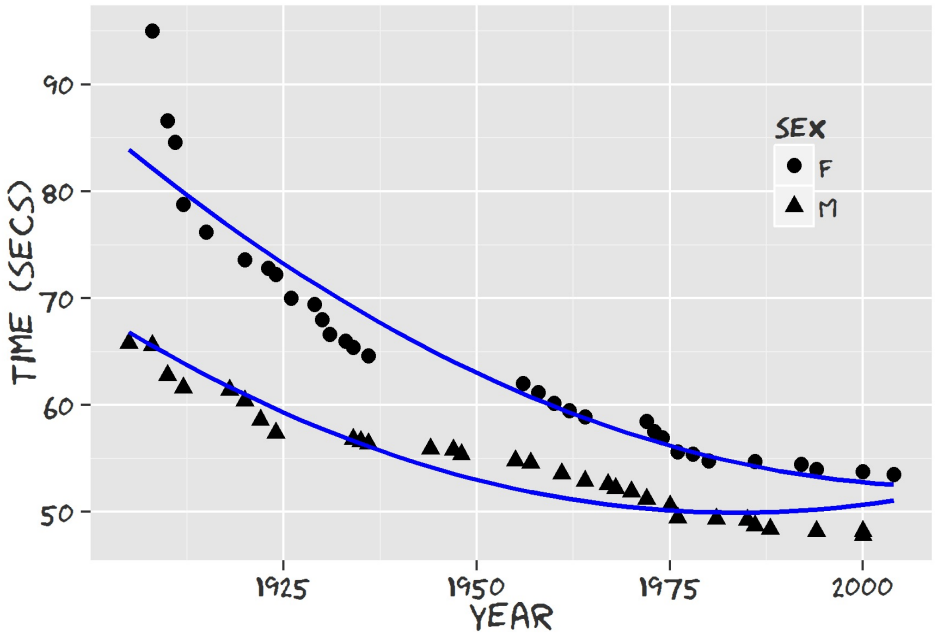
NORMAL BY NATURE

- Random biological or physical processes affected by **large** number of *random processes* with **individually** small effects…
- Sum of all these random components creates random variable that converges on normal distribution
- Regardless of the underlying distribution of processes causing the small effects!

WHAT DOES THIS HAVE TO DO WITH MODELLING?

- The relationships in models are based on sample data – therefore have randomness
- They could change if repeated
- Need to use randomness to evaluate precision of models

- We can use visualization to better understand the characteristics of a given model



MODELS ARE A BIG PART OF R

- Each take some data
- Attempt to generalize about the underling data-generating-process

- Visualization of models is key to understanding what you can do with information embedded within them
  - regression trees e.g.,

GG 501 SPATIAL KNOWLEDGE MOBILIZATION

Feb 8: Model data