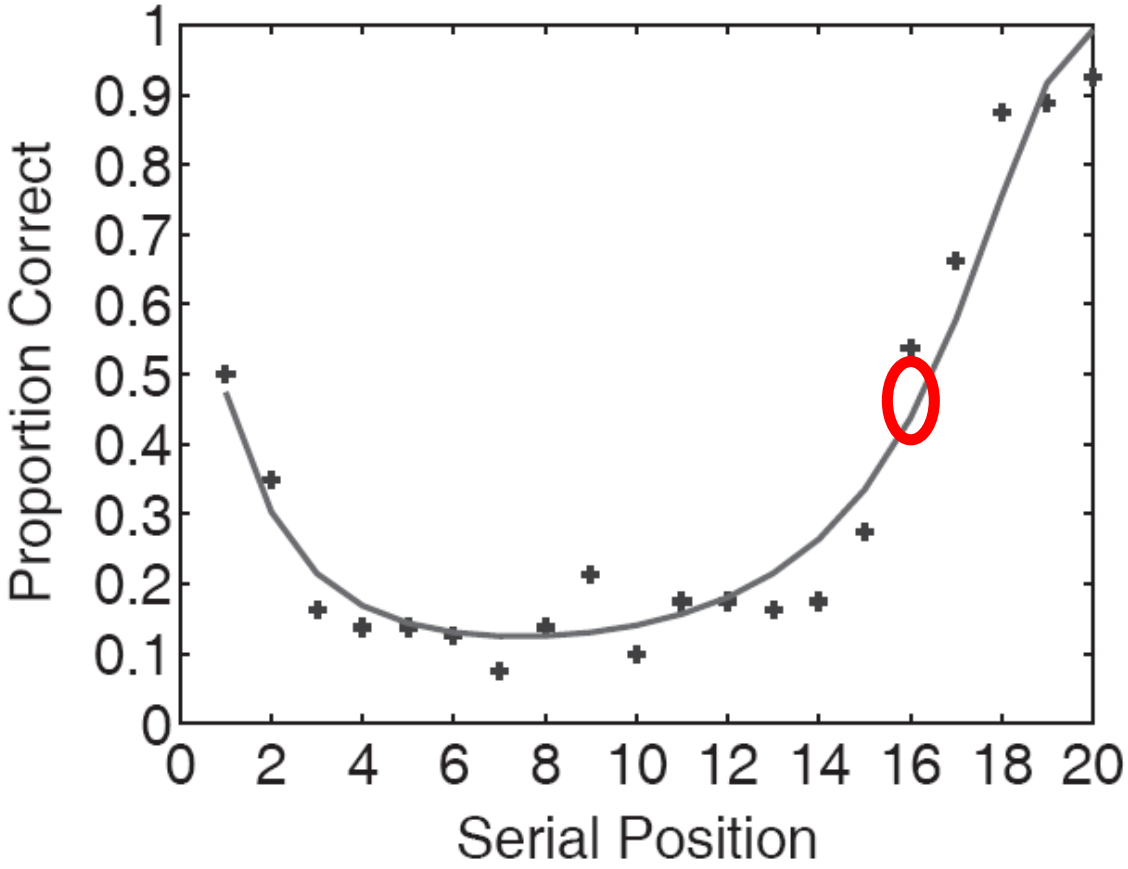
MAXIMUM LIKELIHOOD ESTIMATION

SIMON FARRELL & GORDON D A BROWN

Slack channel #MLE

SIMPLE: fit to free recall data **from a single participant**



What does the predicted probability of 0.5 for serial position 16 mean?

What does predicted probability of 0.5 actually mean

- Does this mean a person is predicted to get exactly 5/10 items correct?
- Would you trust someone in a gambling game whose coin always gave exactly 5 heads for every 10 throws?
- Although the probability in both cases (data and model) is a single value, we will get different results for every set of 10 coin tosses
- Why?
  - ????

- SIMPLE is a deterministic model: each time we run it we get exactly the same predictions (when same parameter values are fed in)
- But people are variable in their responding!

Sampling variability

- In experiments, if we test different samples from a population (or even the same person at different time points) we will get different results
- Why? Sampling variability

Variability from deterministic models

- Just like coin tossing, a predicted probability correct of 0.5 from SIMPLE is a long-run probability
- For a particular set of 10 trials, the person may get 3, 5, or even 10 items correct

Exercise: simulate this process

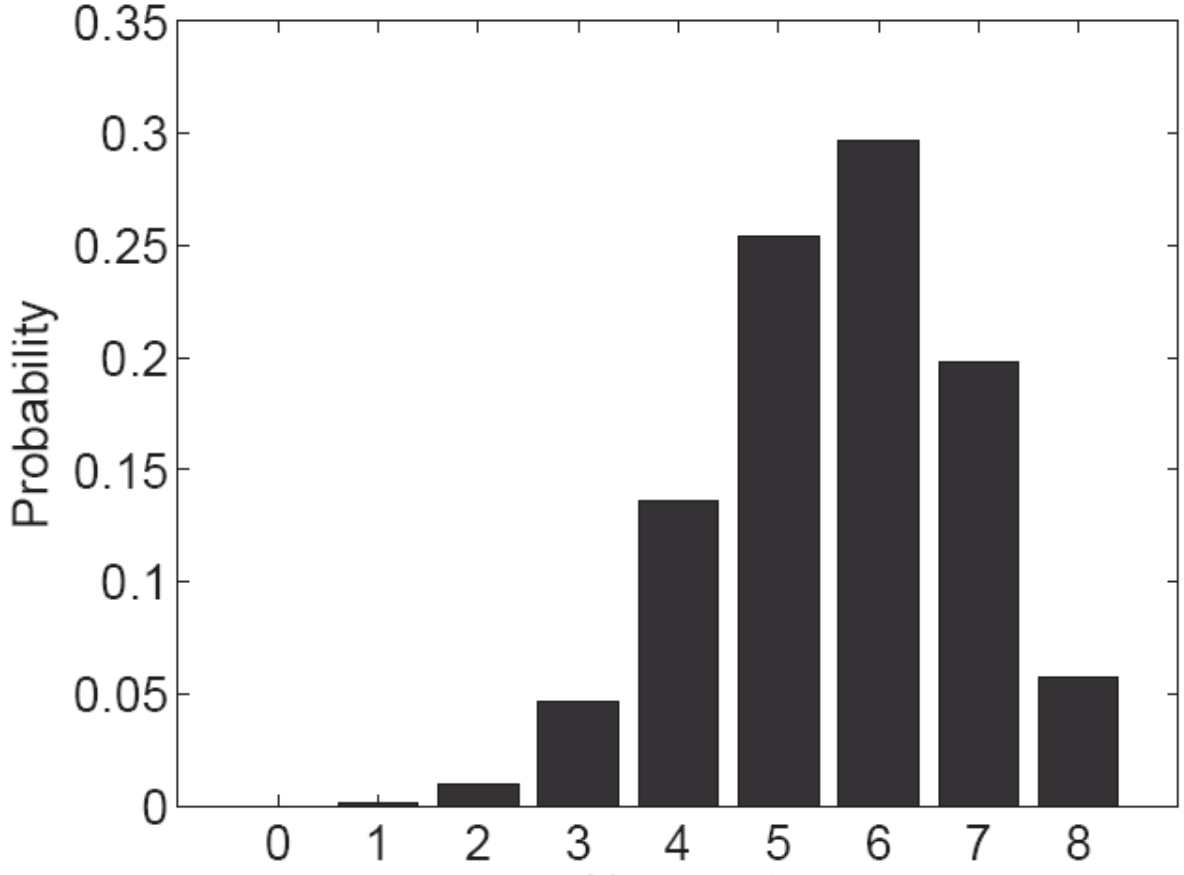
- Monte Carlo simulation of flipping of a weighted coin
  - Probability of heads = 0.7
  - 1000 games
  - For each game, simulate 8 coin tosses and record the observed number of heads
    - A vector of length 1000, each element is the number of heads for a single game
  - Plot a histogram of the number of heads
    - (each score entering the histogram is the number of heads from a single game)

Hint

- Two possibilities
  - runif()<p_heads
  - rbinom

A distribution

- A distribution assigns probabilities to different possible events
- We can describe the distribution we just simulated mathematically
- Rather than simulate, we can work out exactly how many heads we expect given *p_heads* and *n tosses*
- **Binomial distribution**



Probability of heads on a single coin toss= 0.7

Each bar: probability of seeing exactly N heads from 8 coin tosses

p(*k* | *p_heads*, *N*) where k is number of heads

The binomial distribution: flippin' coins

$$p(k|p_{heads}, N) = \binom{N}{k} {p_{heads}}^{k} (1 - p_{heads})^{N-k}$$

"From N choose k"

`choose` in R

- Probability of *k* outcomes actually happening (e.g., getting 5 heads)
  - given *N* total observations (e.g., 8 coin flips)
  - and p_{heads} probability of the event happening on each observation (throwing a head)
- Each *k* has a probability between 0 and 1 (inclusive)

Probability mass function

- The binomial is a probability mass function (also called “probability distribution”)
- Binomial is probability of various **discrete** events given
  - Probability of occurrence on each observation
    - e.g., Getting a head, correctly recalling an item from the study list
  - Number of observations
    - Fixed by the experimenter

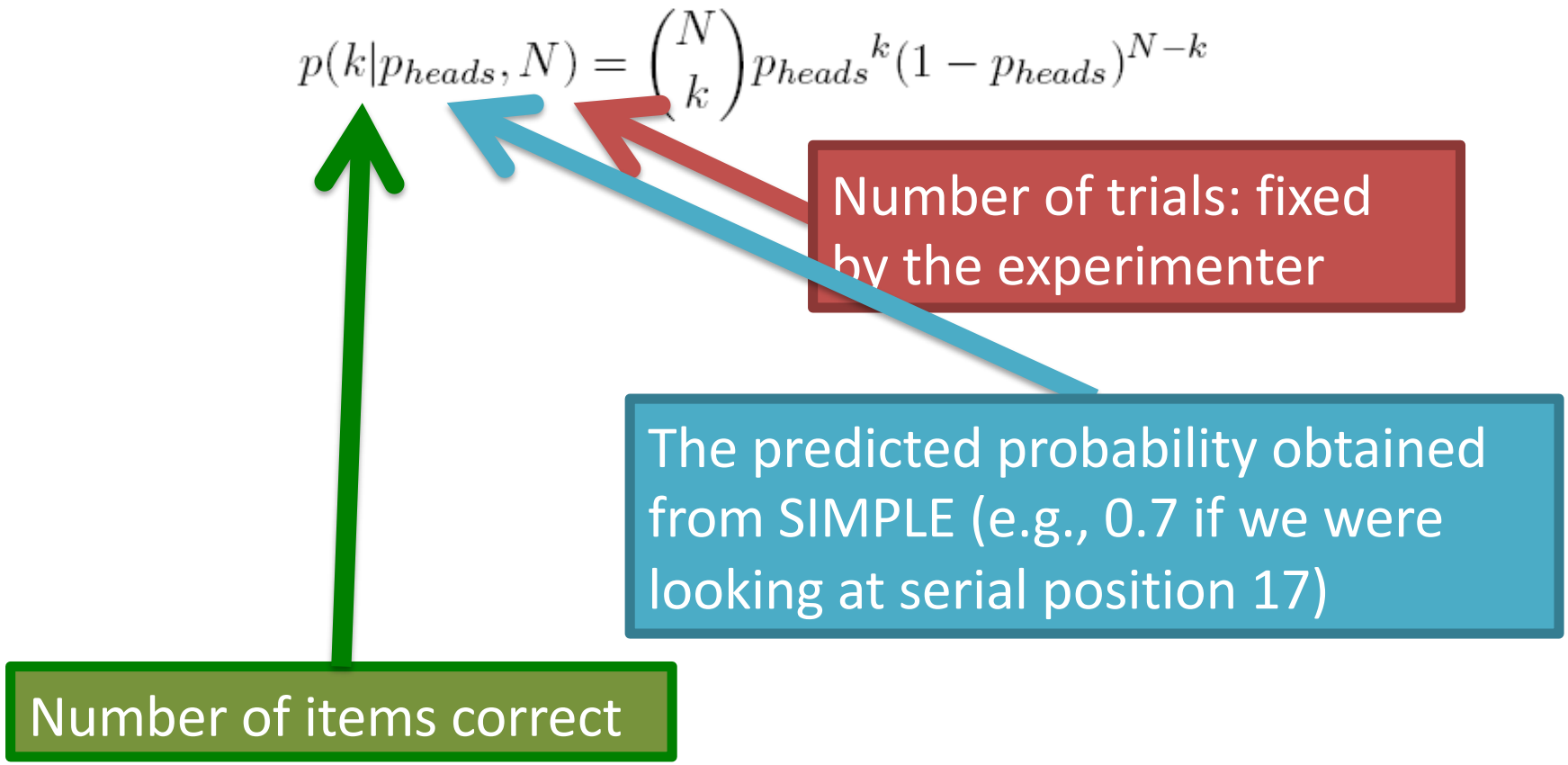
Exercise

- Plot predicted probability distribution: number of heads (0-10) from a coin with p_{heads}=0.5, and with 10 coin tosses in total
- Use the `dbinom` function in R
  - x: the values on the x axis (different possible number of heads)
  - size: total number of tosses
  - prob: the probability of a head
  - Use type="h" when plotting
- Advanced: simulate for p_{heads}=0.7 (N tosses = 8), and plot against the numerical simulation results from earlier

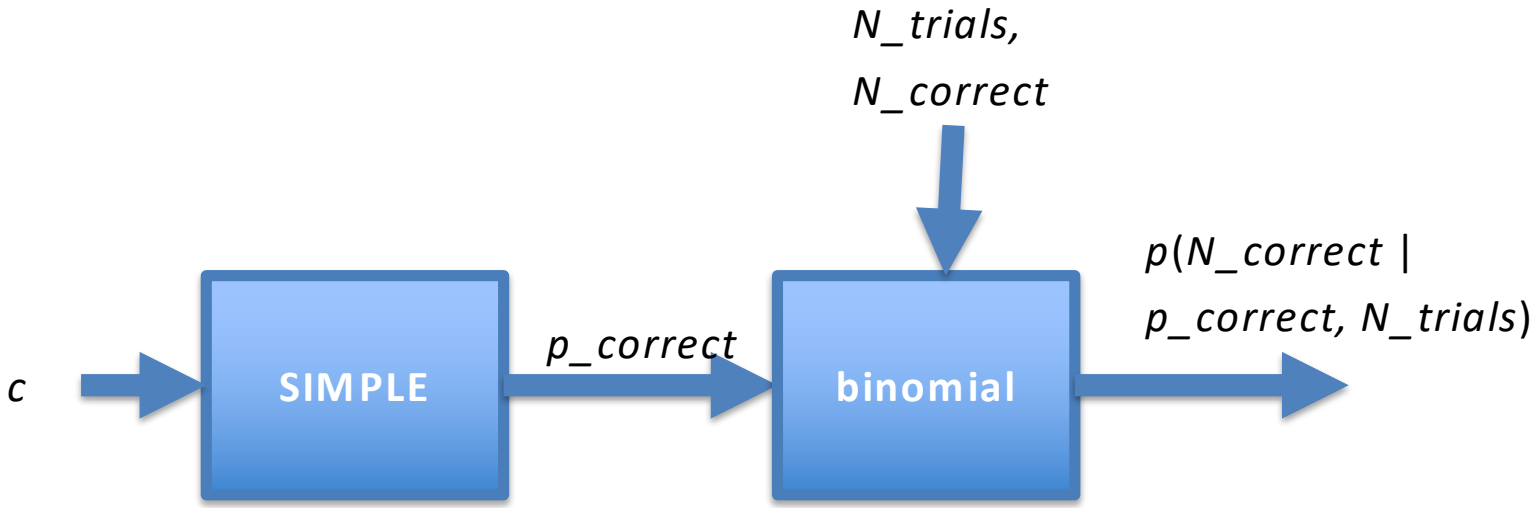
Binomial distribution as a data model

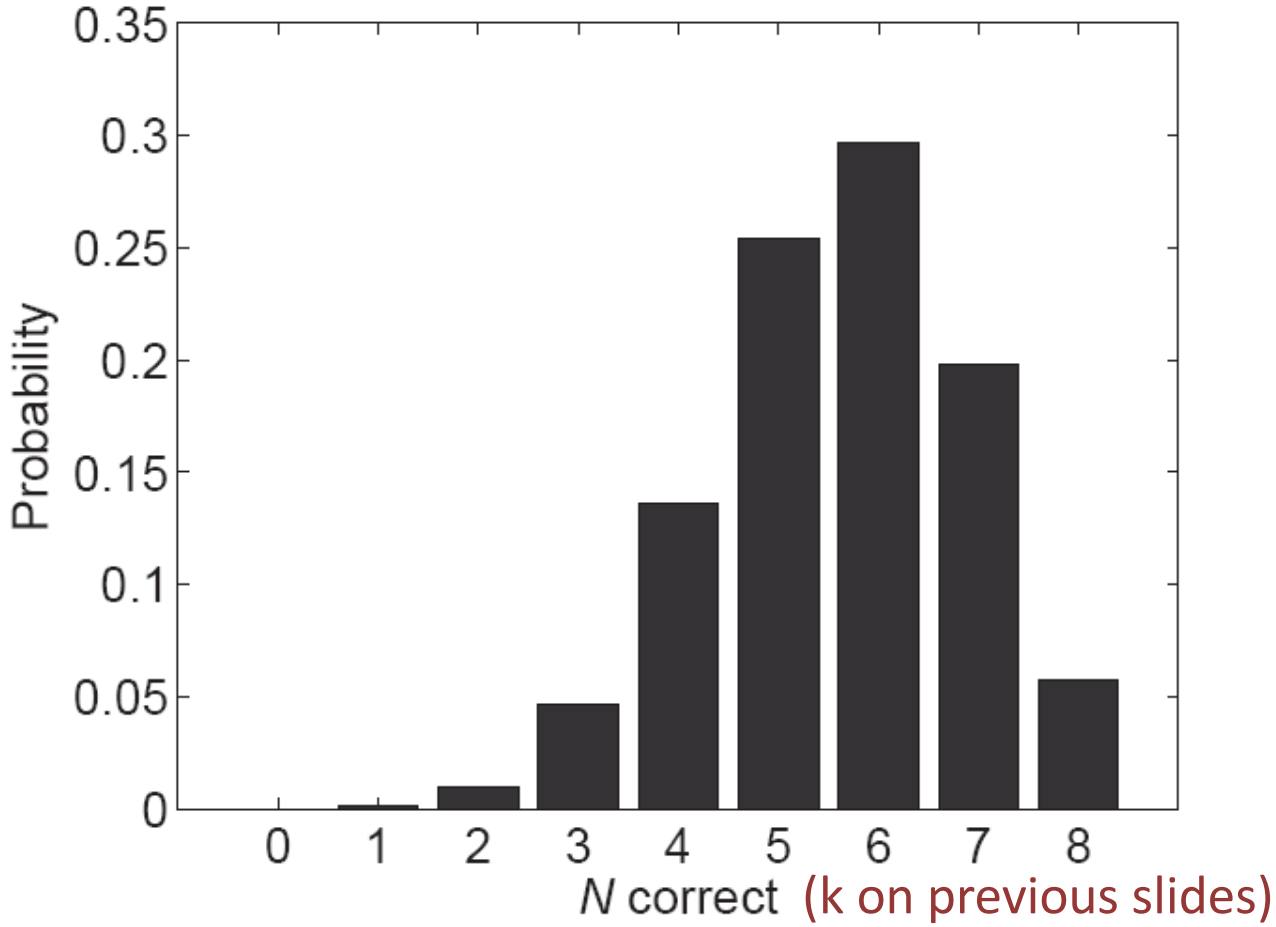
- We can use the binomial is a “data model”
- Allows us to connect model predictions (e.g., predicted probability correct) to empirical observations (number of events, such as number of correct responses)
- Predictions about number of 2-alternative events
  - Number of children passing Sally Ann task (out of, e.g., 10)
  - Number of correct responses at serial position 3 in a free recall experiment for one person (8 trials in total)
  - Number of votes for one of two candidates in an election

How does this all work for SIMPLE (using coin tossing equation)?



Using SIMPLE with binomial





Predicted probability of getting N items correct (for various N correct) from SIMPLE with binomial model

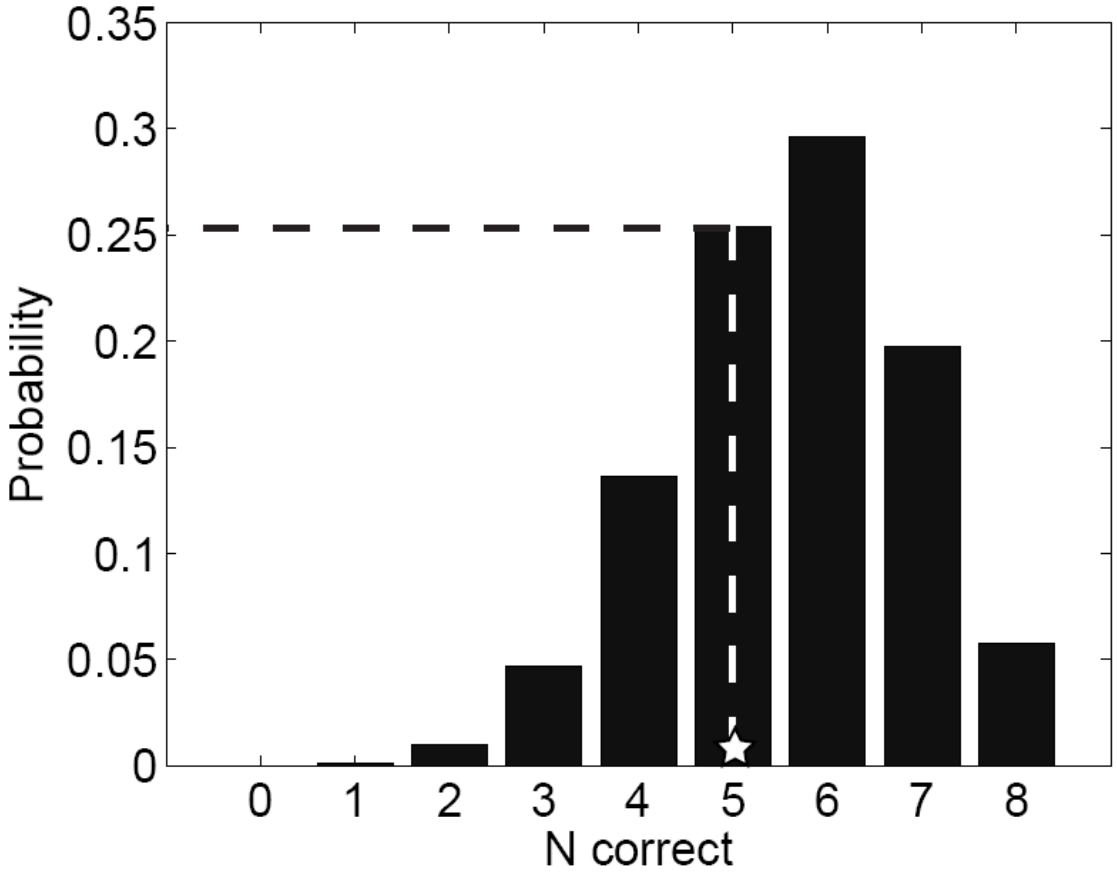
One such distribution for each serial position 😬

Main point so far

- Model prediction is a distribution across possible events (possible data)
- Some models don’t do this, so we need a “data model” to incorporate sampling variability

CONNECTING MODEL TO DATA

- When we run an experiment we only have a single set of data, and a single number correct
  - (for each participant at each serial position)
- How do we connect this to the range of outcomes now predicted by SIMPLE?



Predicted probability correct from SIMPLE = 0.7 (8 trials in total)

Actual data: 5 correct

p(data | p_correct): Probability of the data given the model (binomial) and the predicted probability correct

(again, this is for one serial position)

Exercise (5 mins)

- Use `dbinom` function to calculate p(data | *pcorrect*) for the SIMPLE example
  - Data: 5 items correct
  - *pcorrect*: 0.7
  - N trials = 8
  - “p(getting 5 items correct given the predicted probability of getting an item correct is 0.7, and given that there are 8 trials in total)”

LIKELIHOODS

A subtle problem...

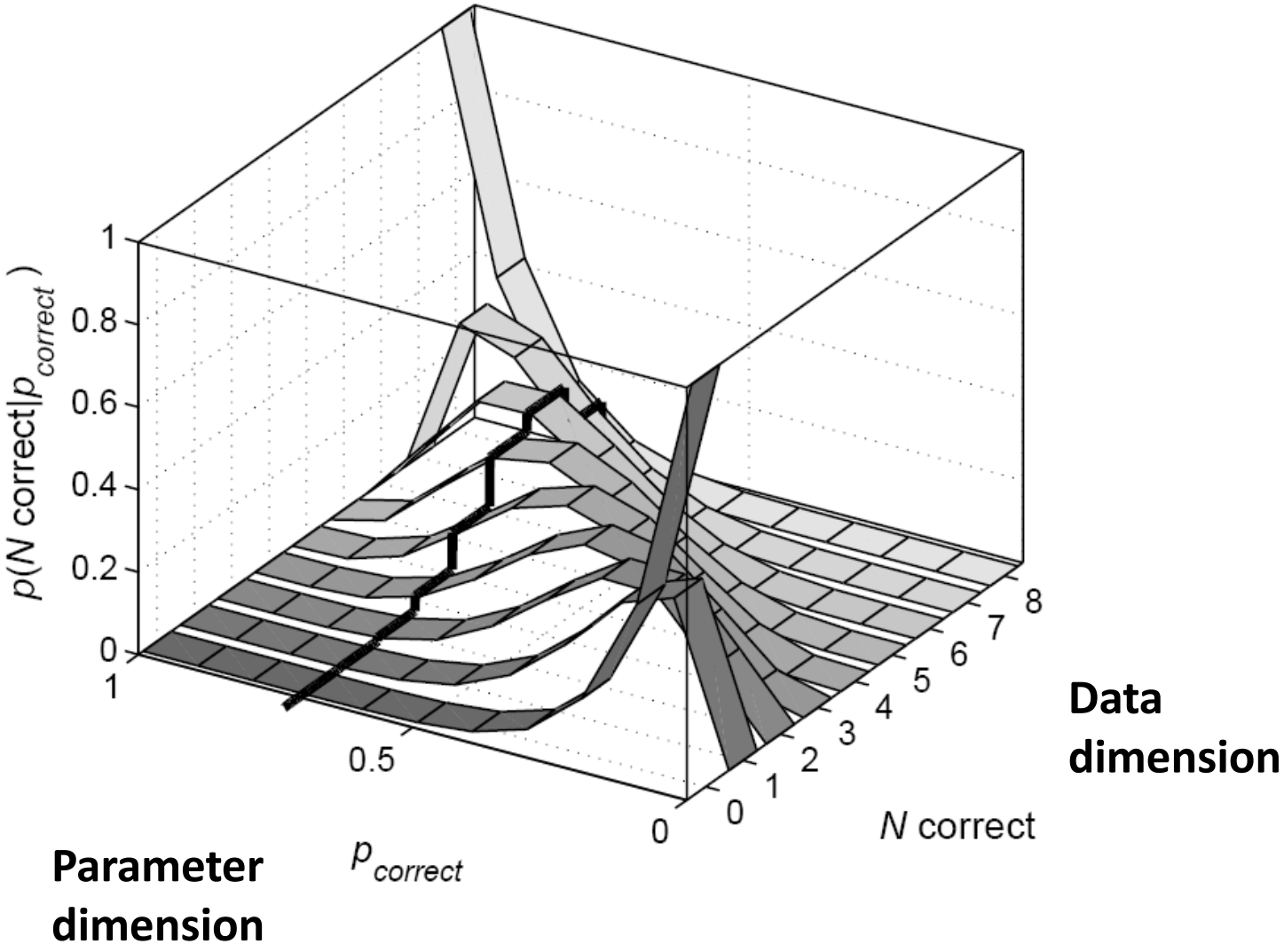
- The binomial distribution gives us the probability of various N correct given the predicted probability
- The predicted probability is determined by the model parameters (e.g., *c* in SIMPLE)
- But...The data are fixed, and we want to estimate the parameters
- **We want to find those parameters that maximize the probability of the data given the model**


I'M SORRY
I'M SO, SO SORRY
quickmeme.com

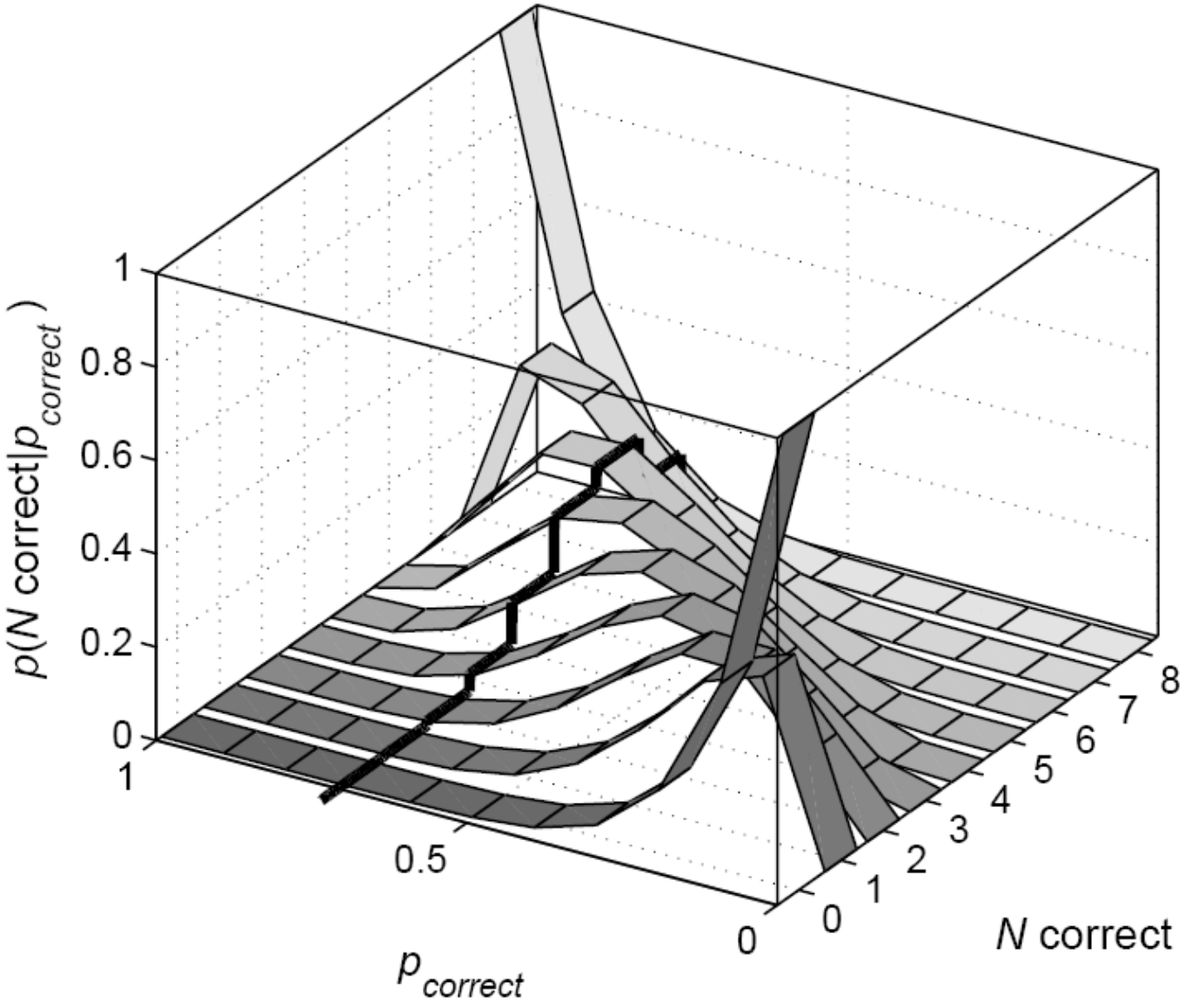
Some geekiness ensues

- Probability function: p(data | parameters)
  - Probability function (e.g. Probability mass function)
- Likelihood function: p(data | parameters)
  - It's the same!
  - But where parameters rather than data change
  - *L*(parameters | data)
  - Give it a different name to reflect the fact that data are fixed, parameters change

- This is not p(parameters| data)
  - Covered in Bayesian modelling

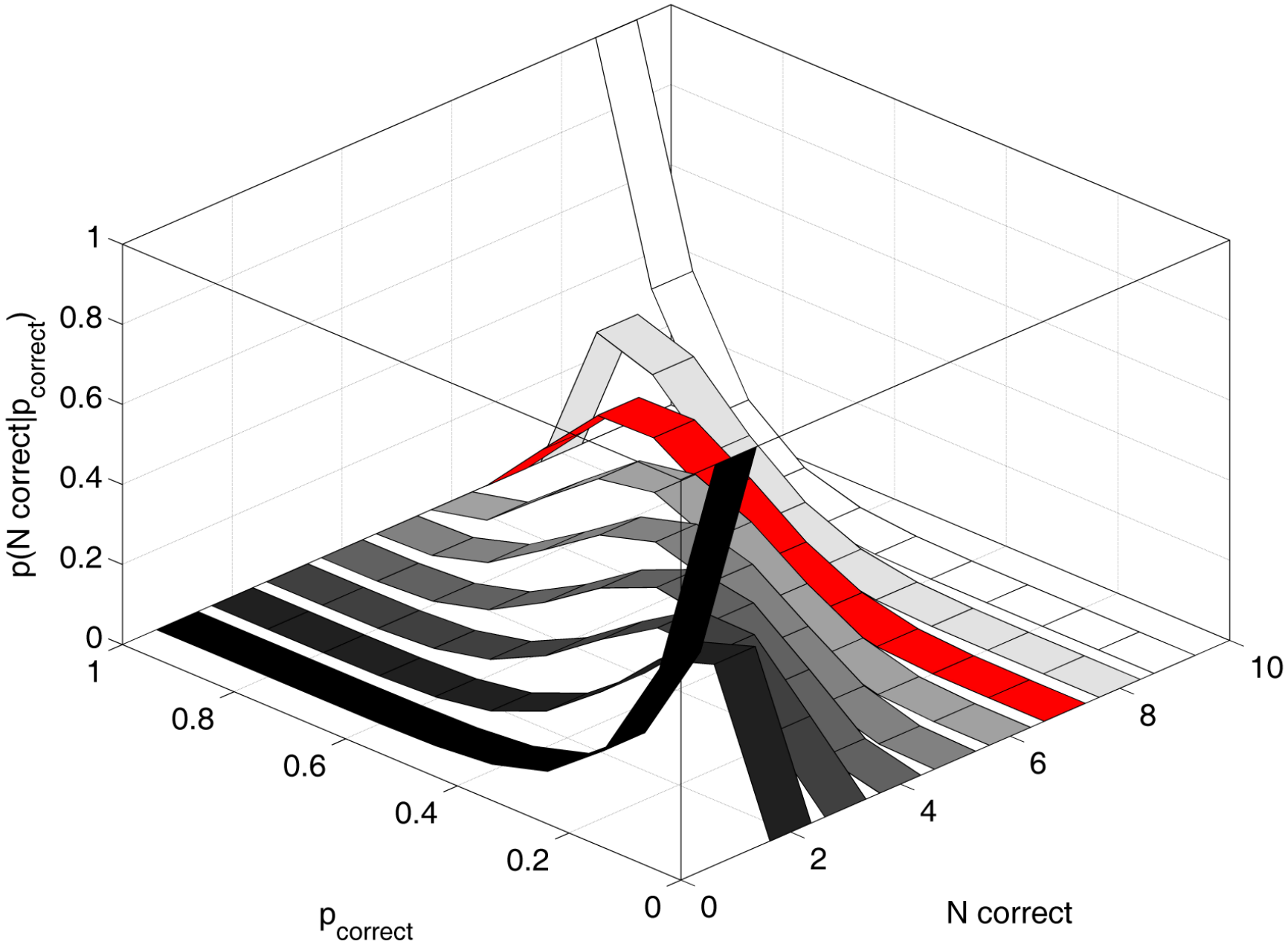


Strips are likelihood functions (continuous)
Dark line is probability mass function (discrete)

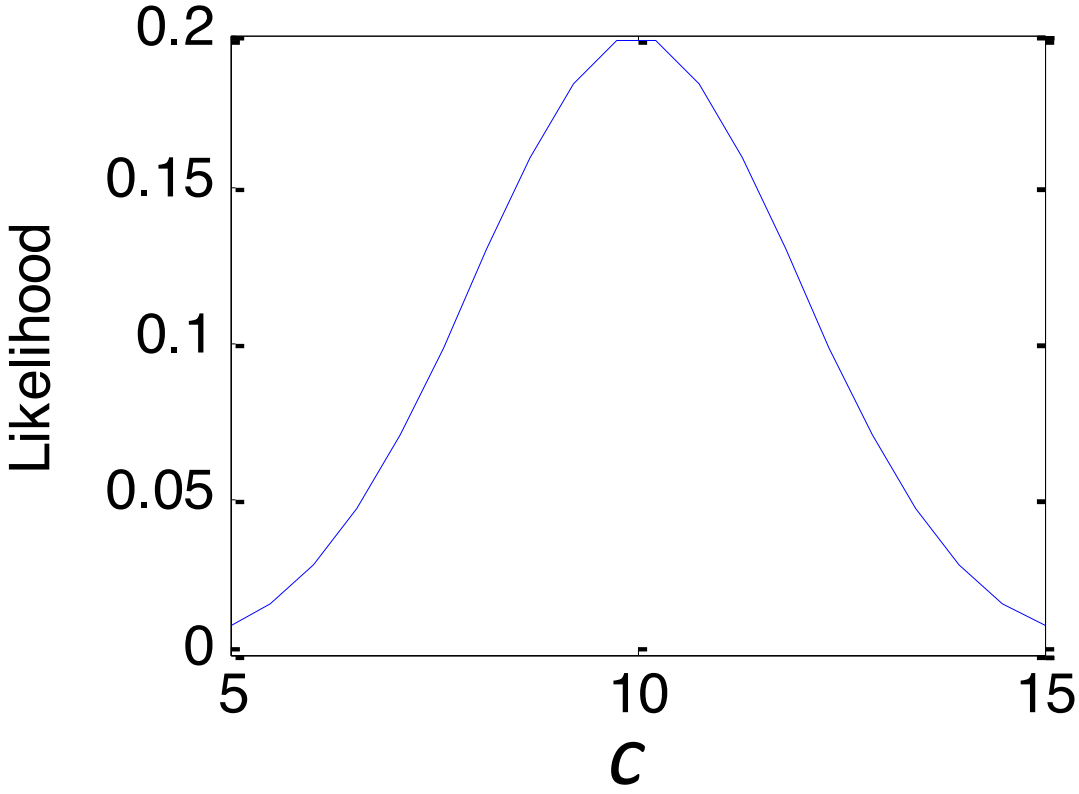


Remember, *pcorrect* is a predicted probability (not model parameters)

Model parameters will map systematically into *pcorrect*


p(N correct|p_correct)
1
0.8
0.6
0.4
0.2
0
1
0.8
0.6
0.4
0.2
0
p_correct
0
2
4
6
8
10
N correct

A likelihood surface



Likelihood function across *c* in SIMPLE
c varies, the data are fixed

MAXIMUM LIKELIHOOD ESTIMATION

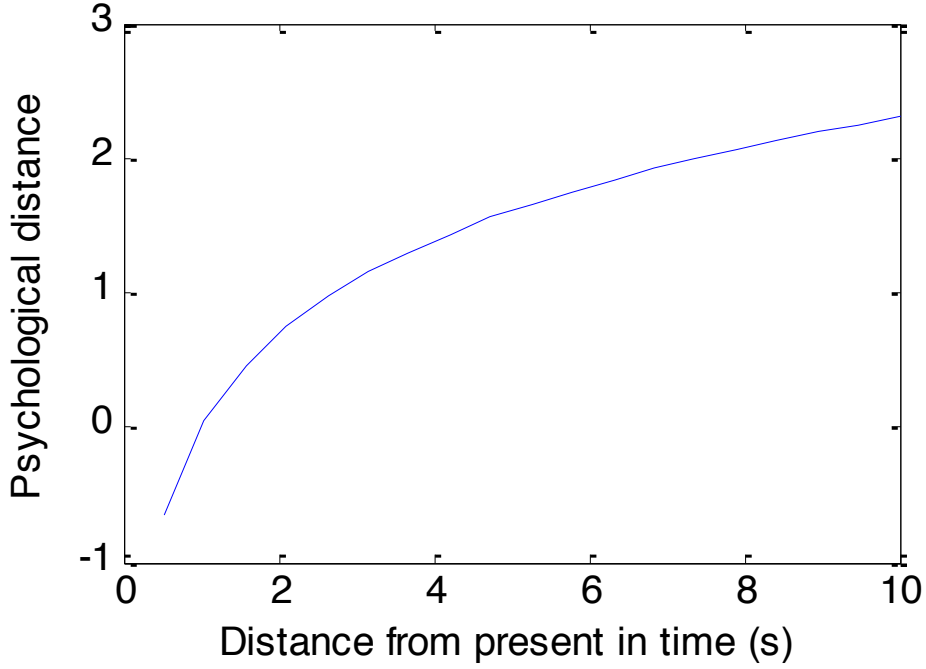
How do we estimate parameters?

- We want to maximize the likelihood
  - Find the peak of the likelihood surface
  - (These can have 2 or more dimensions)
- Or minimize the negative likelihood
  - Remember, optim() does function **minimization**
- Can do this using methods from earlier today (SIMPLEX)
- But first…

A detour through log space

Convention is to work with log-likelihoods

- log(x)
- Compression
- SIMPLE: temporal compression
- Orders of magnitude on linear scale
  - Log10 scale:
  - 1-10-100-1000 on x
    1-2-3-4 on y



Natural log 

- Natural logarithm (ln): inverse of exponential
  - exp(x) = e^x (e = 2.7183)
  - $e^{x:}$ 1 2 3 on x maps on to e^1 e^2 e^3 on y
  - $\log_e$ scale: e^1 e^2 e^3 on x maps on to 1 2 3 on y

Log-likelihood

- Work with log-likelihood function (ln *L*) rather than likelihood function
  - Natural logs
- Makes the job easier: numbers are smaller and less likely to go out of range of computer
- **Log likelihoods add up**

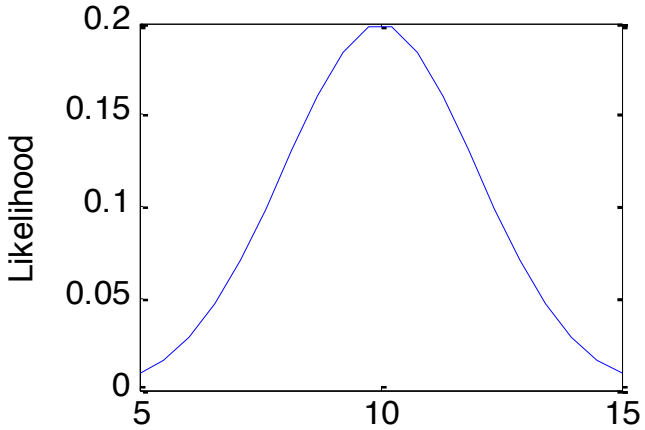
Log-likelihood is a principled measure of fit

- -2 ln *L*: *deviance*
- Related to chi-square that Steve talked about earlier (briefly)
- Statistical measure of discrepancy between model and data (or “reality”)
- As we’ll see later in the school, deviance can be used to compare fit of different models

Maximum likelihood estimation

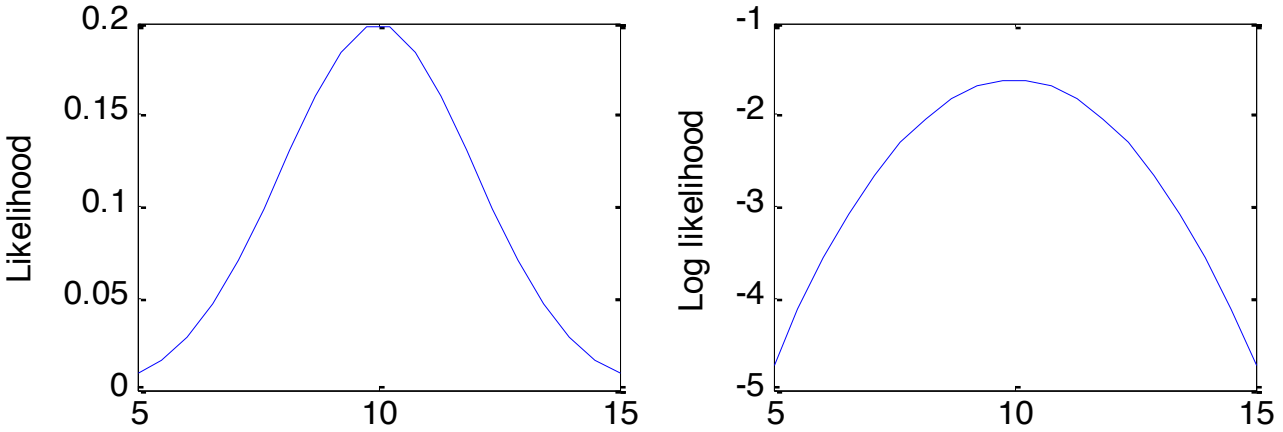
- Find parameters that maximize the likelihood
- Find those parameters that maximize the probability of the data given the parameters
- In practice: minimize negative log-likelihood
  - Allows us to use SIMPLEX etc.
  - Double the minimized negative log-likelihood to get deviance

Likelihoods are confusing

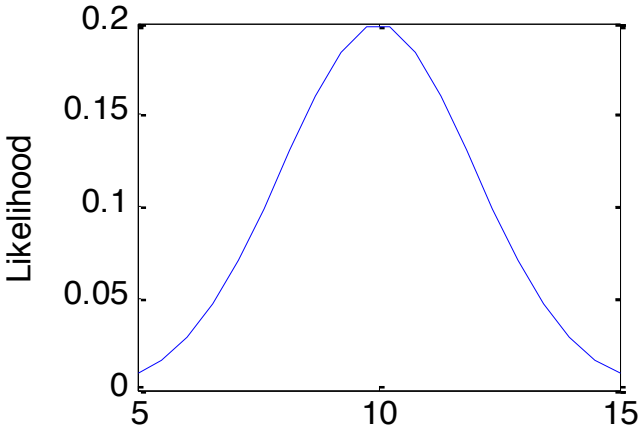


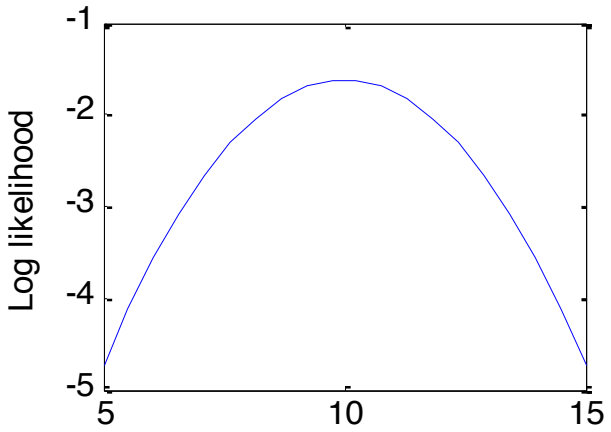
Likelihood function across c in SIMPLE

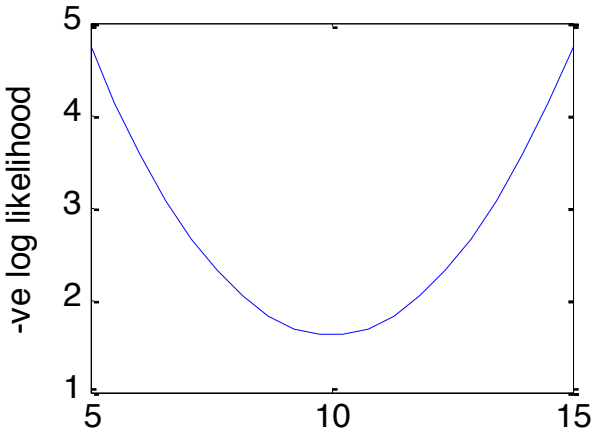
Likelihoods are confusing



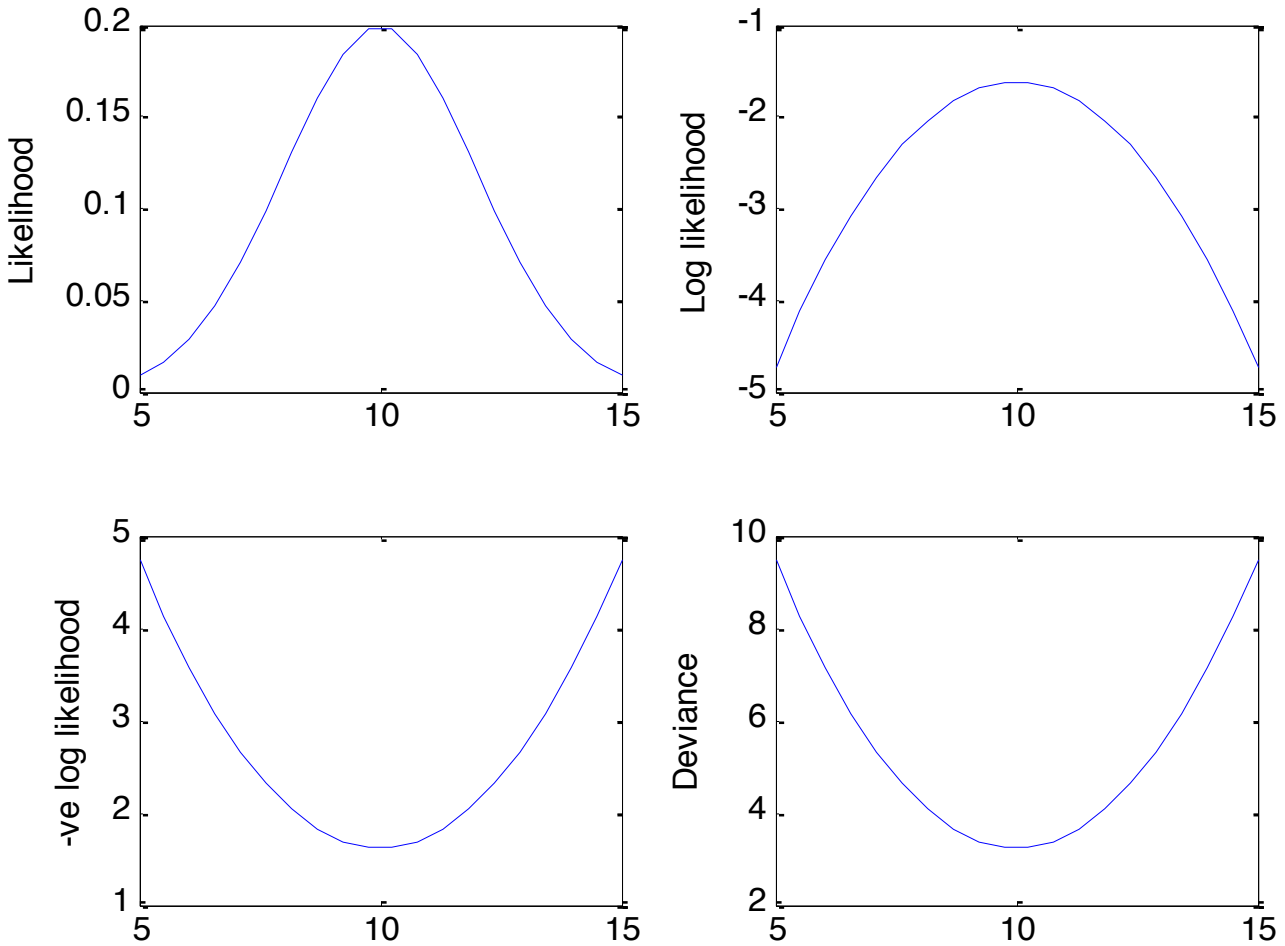
Likelihoods are confusing







Likelihoods are confusing



OVER TO GORDON

- Number correct/passed/choice from two alternatives: binomial
- More than two categories: multinomial
  - Serial recall
    - Correct
    - Order error (list item recalled in wrong position)
    - Item error (non-list item recalled)
- What about average proportion correct, or variables like RT?

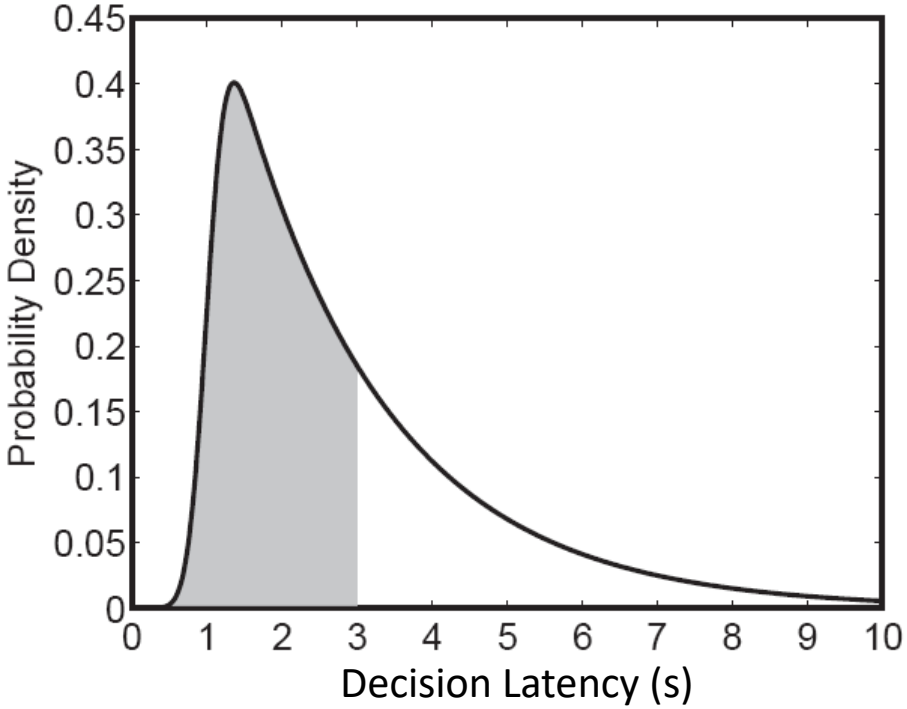
Continuous distributions

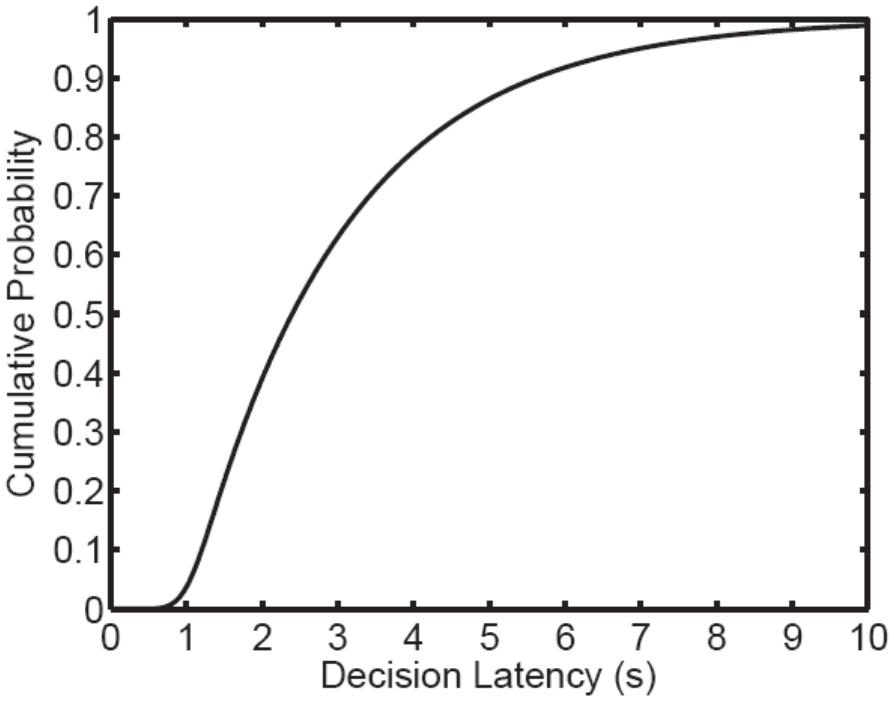
- We don’t have discrete outcomes for a continuous distribution
  - Effectively infinite number of possibilities
- Each possibility effectively has 0 probability
- Instead, talk about probability **density**

LIKELIHOODS FOR CONTINUOUS DISTRUBITIONS

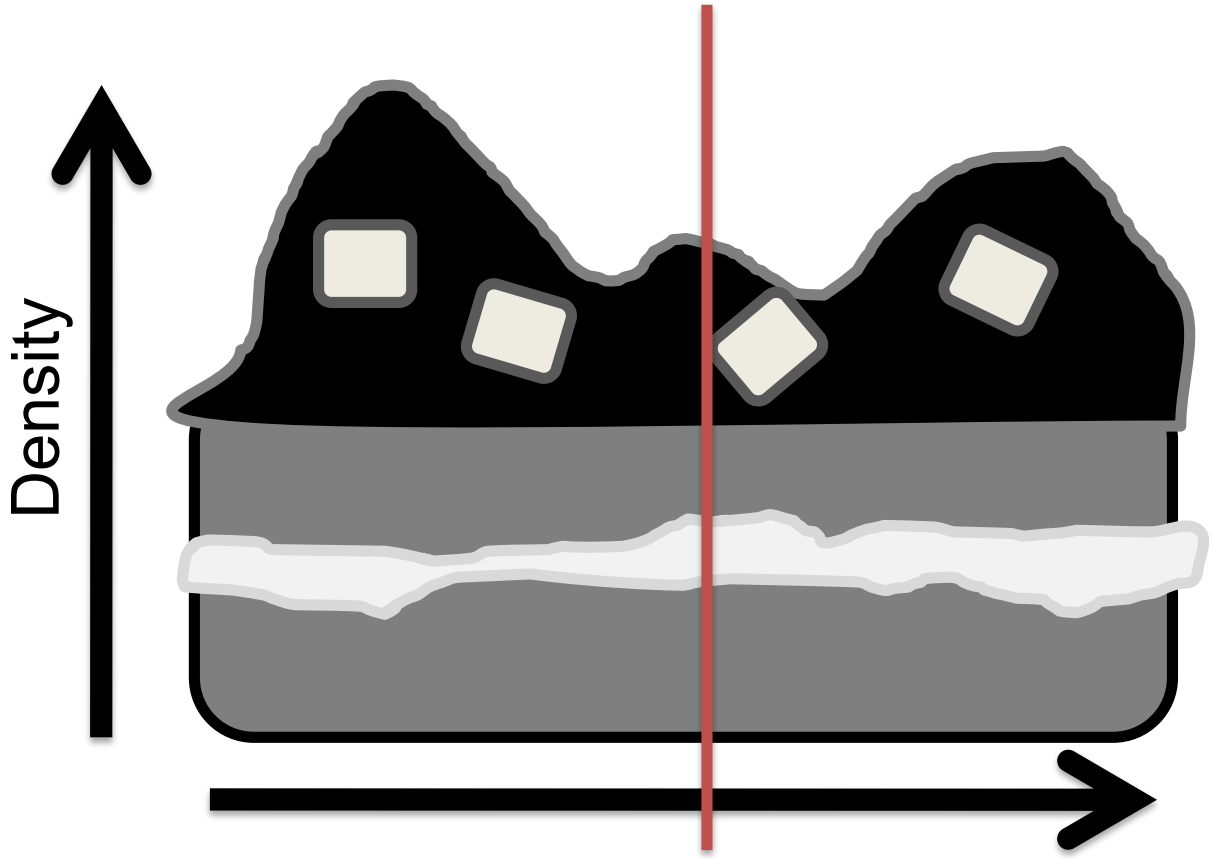
Continuous distributions defined by probability density

- Can't assign probabilities to discrete categories—there aren't any!
- Instead, refer to **density** of curve

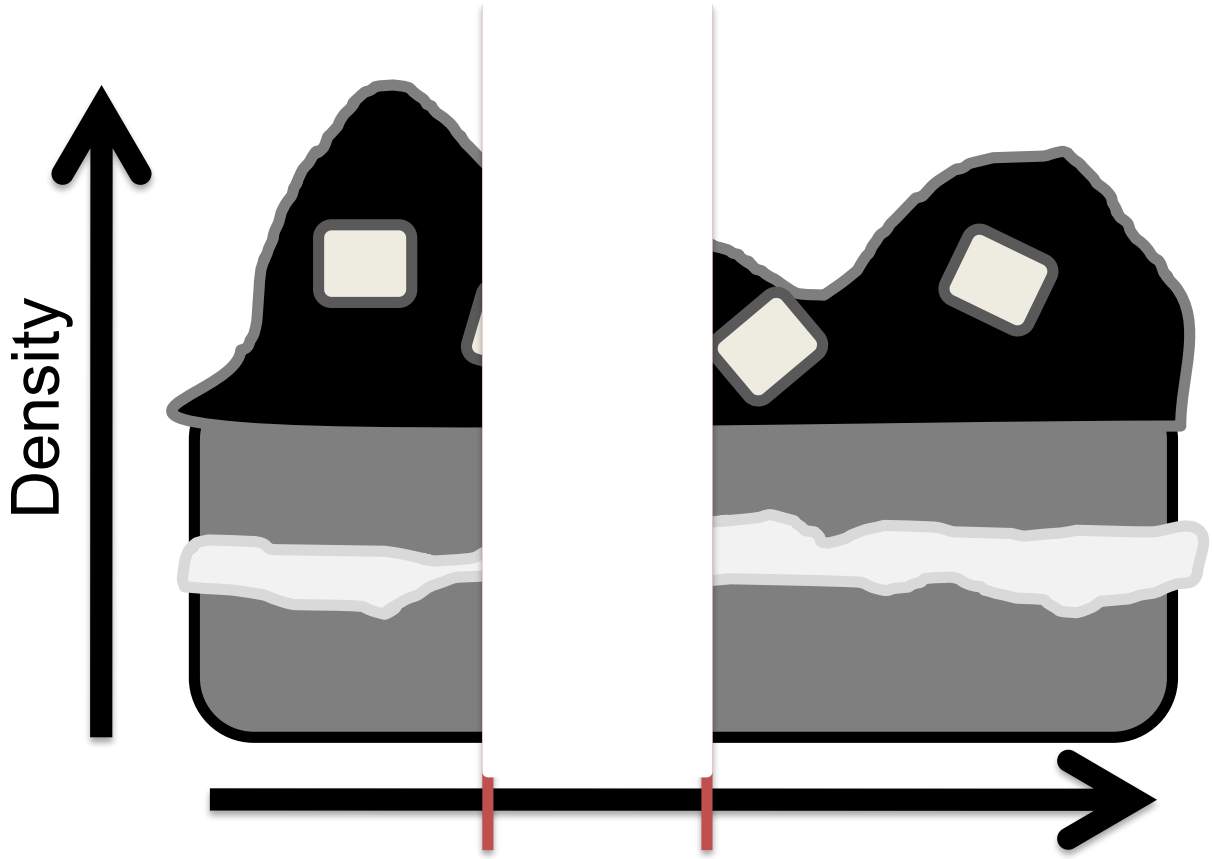




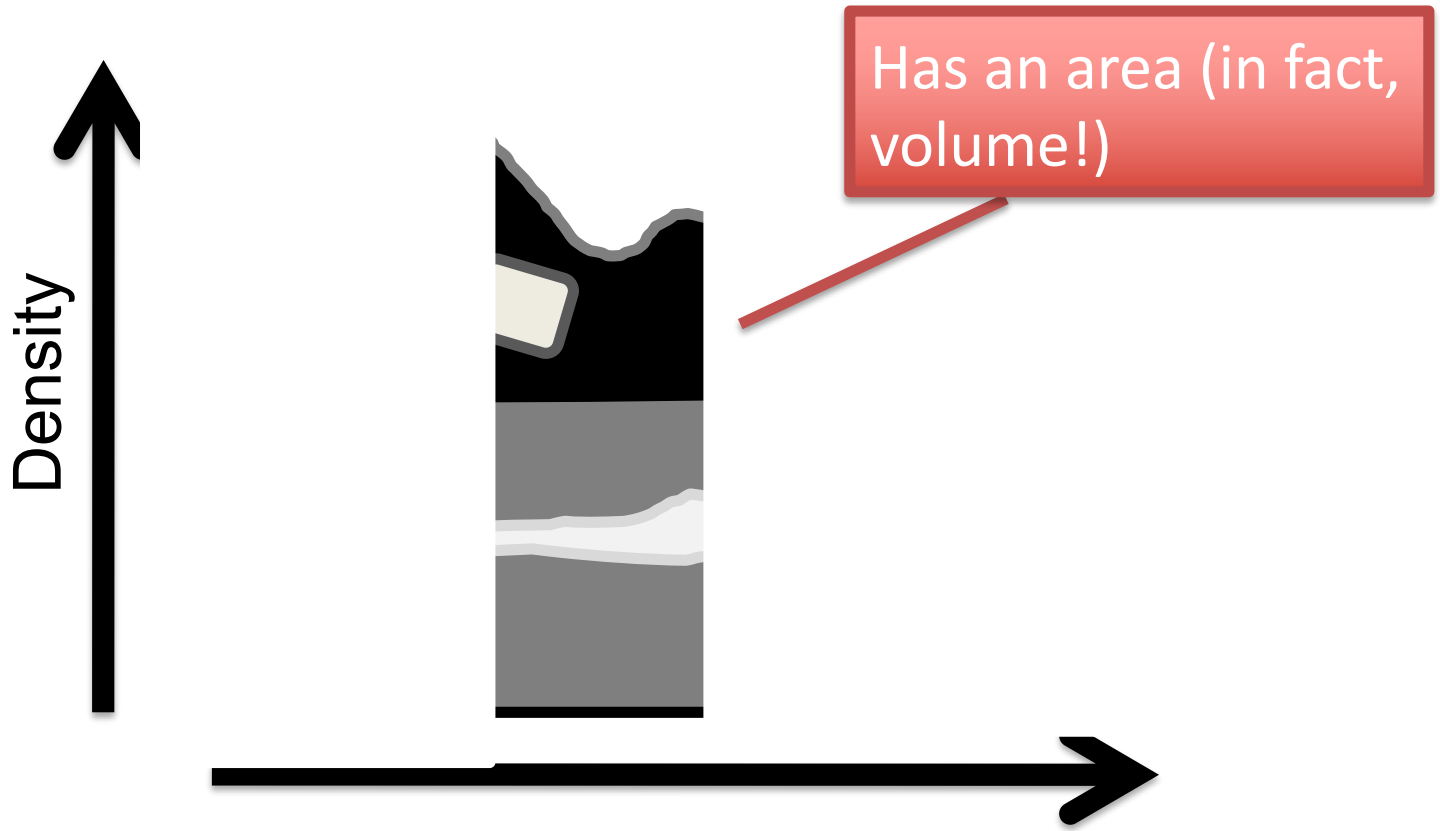
Cake space



Cake space



Cake space



We've used some probability density functions already on previous days

- ???

Example: the shifted Weibull

$$f(x) = \left(\frac{\beta}{\theta}\right)\left(\frac{t-\psi}{\theta}\right)^{\beta-1} \exp\left[-\left(\frac{t-\psi}{\theta}\right)^{\beta}\right]$$

- Used to model response times
  - E.g., Cousineau et al. (2004); Rouder et al. (2004)
  - ψ(psi): shift
  - θ (theta): scale
  - β (beta): shape

Exercises

- 1. Plot probability density function from the Weibull for scale = 200, shape = 2
  - `dweibull`
  - (We will assume shift=0 for the moment)
  - Across the range 0-1000 ms
- 2. What is p(data|parameters) for data = 200 ms and the given parameters?

Exercises 2

- Read in the 200 RTs from rt.txt
- Fit the Weibull to the data using maximum likelihood estimation

Maximum likelihood estimation with the weibull

- Things you will need to be doing
  - Calculate log likelihood for a single data point under Weibull
  - Extend this to calculation for multiple data points
  - Convert to – ln*L* and sum
  - Wrap in a function that takes two arguments: theta (vector of parameters) and data vector
  - Fit to data using optim()

Summary: likelihoods

- Predictions are distributions across data space
- Fundamentally grounded in statistical theory
- Recognize variability/uncertainty in behaviour
- A key ingredient in Bayesian modelling
- Allow quantitative comparison of model fits (AIC/BIC, later in the school)

END