Likelihood and Bayes' Theorem

Eric Scott

2020-1-23

1. Short problem from the end of last class

1. Ignoring flowering (for now), explore the likelihood of different values of survival, given the 5-plant data set.

   a. Calculate the likelihood of survival having each of the following values: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

```
s <- seq(0.1, 0.9, 0.1)
s
```

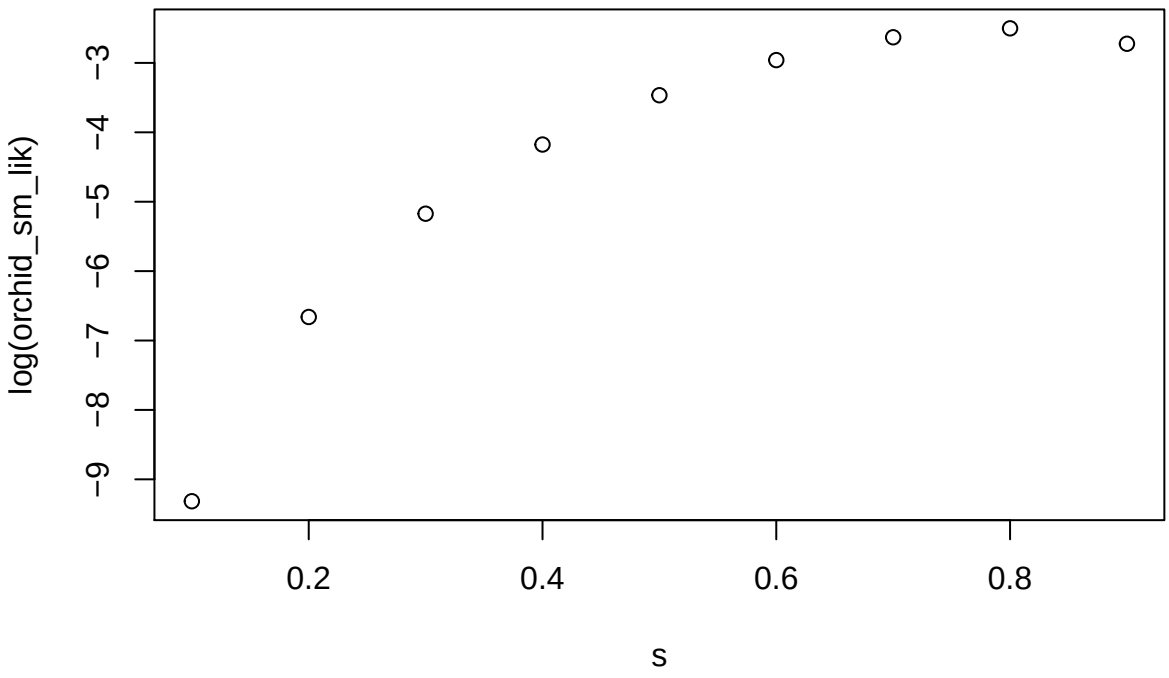
```
## [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
```

```
orchid_sm_lik <- s^4 * (1-s) #vectorized over all values of s
orchid_sm_lik
```

```
## [1] 0.00009 0.00128 0.00567 0.01536 0.03125 0.05184 0.07203 0.08192 0.06561
```

   b. Make a graph of the log-likelihood (y-axis) vs. value of survival (x-axis).

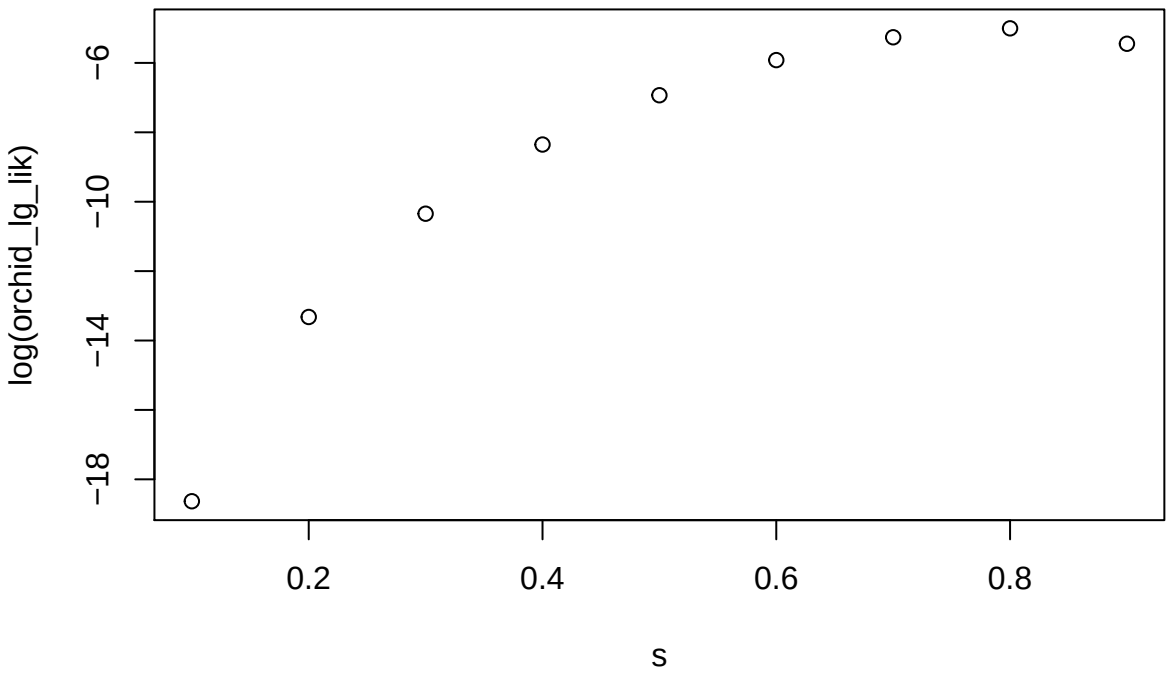
```
#base R plotting
plot(s, log(orchid_sm_lik))
```



c. Repeat a&b for the 10-plant data set. How does the shape of the graph change?

```
orchid_lg_lik <- s^8 * (1-s)^2
```

```
#base R plotting
plot(s, log(orchid_lg_lik))
```



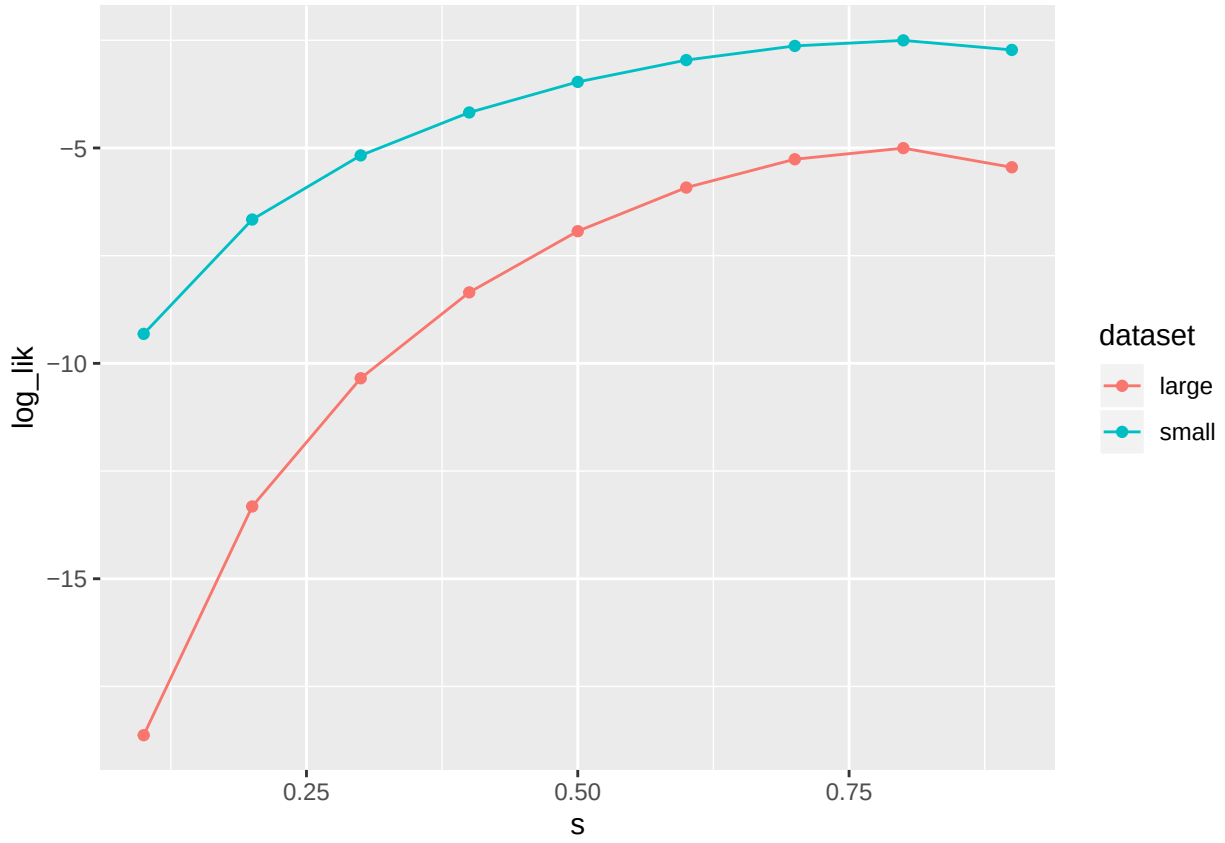

```
#using ggplot2 to put both lines on the same plot
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------------ tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# make a 'tidy' dataframe/tibble for plotting
df <-
  tibble(
    dataset = c(rep("small", 9), rep("large", 9)), #column indicating which dataset
    s = c(s,s), #x-axis values
    likelihood = c(orchid_sm_lik, orchid_lg_lik),
    log_lik = log(likelihood)
  ) #you can do math while creating a tibble

head(df)
```

```
## # A tibble: 6 x 4
##   dataset     s likelihood log_lik
##   <chr>   <dbl>      <dbl>   <dbl>
## 1 small     0.1    0.00009   -9.32
## 2 small     0.2    0.00128   -6.66
## 3 small     0.3    0.00567   -5.17
## 4 small     0.4    0.0154    -4.18
## 5 small     0.5    0.0312    -3.47
## 6 small     0.6    0.0518    -2.96
```

```
ggplot(df, aes(x = s, y = log_lik, color = dataset)) +
  geom_point() +
  geom_line() #add connecting lines
```



d. What is the likelihood that survival is 0? Or 1? What happens to the log-likelihood at these values?

```
0^8 * (1-0)^2
```

```
## [1] 0
```

```
log(0)
```

```
## [1] -Inf
```

2. Thinking about probabilities...

a. What is $P(B|A)$ for two mutually exclusive events? Is it possible for events to be independent and mutually exclusive?

$P(B|A) = 0$ because there is no overlap. Not possible for events to be independent and mutally exclusive. E.g. heads & tails aren't independent.

b. What is the probability that A **or** B occurs if they are **not** mutually exclusive?

$P(A + B) = P(A) + P(B) - P(A, B)$

2. Homework #1: events, scope of inference, sample size & estimated probability

4. Introduction Bayes' Theorem (as time permits)

So far three major conceptual points:

1. Using data to estimate parameters (in this case, the survival and flowering probabilities of plants)
2. Calculating the probability of collecting a particular data set if a particular model were true ($P(data|model)$)
3. Calculating *likelihood* (relative support) for a model, given a data set

Now:

4. Calculating the probability a model is "true", given a data set

Bayes' theorem is an alternative method of assessing the **probability** of a model given a fixed set of data.

Rev. Thomas Bayes (1702-1761) showed that we can convert the probability of seeing a particular data set, given a model and parameters, into the probability of the model and parameters, given the data:

Bayes Theorem:

Recall the definition of conditional probability:

$$P(B|A) = \frac{P(A, B)}{P(A)}$$

because it's symmetric

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

rearranging the 2nd equation:

$$P(A, B) = P(A|B) \times P(B)$$

substituting this into P(A,B) in the first equation:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$$

If B is the 'model' and A is the 'data', this is Bayes' Theorem.

$$P(model|data) = \frac{P(data|model)P(model)}{P(data)}$$

$P(model|data)$ = probability of model and parameters, given the data $P(data)$ = constant specific to the data set $P(model)$ = sometimes controversial "prior probability"

Simple intuitive example: testing for Lyme Disease

- We have tests for Lyme disease, so why shouldn't everyone get tested regularly?
  - true positive rate = 87%
  - false positive rate = 1%
  - (these numbers actually vary depending on the type of test)
- But Lyme Disease is pretty rare in the US. According to the CDC, about 300,000 cases per year
- 300,000 / 300,000,000 = 0.001, or about 0.1% of population

$$P(Lyme|+) = \frac{P(+|Lyme) \times P(Lyme)}{P(+)}$$

Or in English:

The probabilty of having Lyme given a positive test equals the true positive rate of the test times the probability of getting Lyme divided by the probability of getting a positive test result.

You can also look at it like the probability of getting a true positive (this model) out of all the ways of getting a positive (both models) adjusted for the actual prevalence of Lyme (your prior belief of how likely it is that anyone would have Lyme).

Defining the terms:

- The "data" in this case, is a positive test result
- The "model" or "hypothesis" is "infected with Lyme disease". There are two possible, mutually exclusive models: has lyme disease, or doesn't have lyme disease.
- In this case, $P(data|model)$ is given to us. But we've calculated this before (we'll come back to the orchid example).

Defining the terms:

- $P(model) = P(Lyme)$ is our prior probability, or $P(H)$ before considering the data. Since we know Lyme is rare, we can use this prior belief to inform the probability of having Lyme given a positive test result.
- $P(data) = P(+)$ All possibilities that match the data. If the models are mutually exclusive, AND we have included all possible models in our analysis, then $P(data) = P(data|model) \times P(model)$ summed over all (mutually exclusive) models. ["OR" axiom of probability]
- In Bayesian analysis, we typically ASSUME all possible models are considered in the finite set of j models we are comparing, and therefore substitute
- In this case, there are only two models, so this calculation is simple.

Calculations

$$P(Lyme|+) = \frac{P(+|Lyme) \times P(Lyme)}{P(+|Lyme)P(Lyme) + P(+|uninfected)P(uninfected)}$$

$P(+|Lyme)$ is the true positive rate and $P(+|uninfected)$ is the false positive rate.

$$P(Lyme|+) = \frac{0.87 \times 0.001}{0.87 \times 0.001 + 0.01 \times 0.999}$$

```
p_lyme_pos = (0.87 * 0.001) / ((0.87 * 0.001) + (0.01 * 0.999))
p_lyme_pos
```

```
## [1] 0.0801105
```

Only an 8% chance you have Lyme given a positive result!!

What would change if you knew the patient lived in Massachussetts? 87,000 cases / 6.5 million = 0.013

```
(0.87 * 0.013) / ((0.87 * 0.013) + (0.1 * 0.987))
```

```
## [1] 0.1028088
```

What if they were bitten by a tick AND had a bulls-eye rash?

```
(0.87 * 0.5) / ((0.87 * 0.5) + (0.1 * 0.5))
```

```
## [1] 0.8969072
```

More on $P(model)$, the "prior probability"?

- The probability of each possible model prior to collection or analysis of the data being analyzed.
- **Informative priors** mean that all models are not equally likely, prior to collecting the data that will be explicitly incorporated into a statistical analysis. Developing informative prior probabilities is a large, active, controversial and mathematically and computationally dense field of research.
- In many cases, we want to base inference only on our data, in which case, if we compare j possible models, the prior probability of each is usually therefore $= 1/j$. This is called an **uninformative prior**.

More on $P(data)$

If the models are mutually exclusive, AND we have included all possible models in our analysis, then P(data) = P(data|model)P(model) ... summed over all (mutually exclusive) models. ["OR" axiom of probability]

In Bayesian analysis, we typically ASSUME all possible models are considered in the finite set of j models we are comparing, and therefore substitute

$$P(data) = \sum_{j=1}^{n} P(data|model_j) \times P(model_j)$$

- This definition of $P(data)$ means that the sum of the $P(model_j)$ over all j models must be 1.
- This definition also makes $P(data)$ constant for a given data set, so, from a practical perspective, Bayes' $P(model)$ begins to resemble Edwards' definition of **likelihood** when there are **uninformative priors**:

$$P(model|data) = \frac{P(data|model) \times (1/j)}{c}$$

- $1/j$ is the [uninformative prior] probability of the model - $P(data)$ is a constant, c, that is the sum of the $P(data|model) \times (1/j)$ over "all" models - note that because $P(data|model)$ is in the equation it is specific to the particular data set(!) - $L(parameters(\theta)|data) = P(data|model, parameters) \times k$ - These are identical if $k = 1/jc$

More on $P(model|data)$

- Also called "posterior probability". The probability of the model *after* seeing the evidence (as opposed to "prior probability")

In other words:

- In Bayesian analysis, we explicitly estimate the constant relating the probability of the model, given the data. This means we can obtain an **absolute** measure of support for a model **IF** we are willing to believe we have searched all possible models.
- In likelihood analysis, we simply use the fact that the likelihood of a model given the data is proportional to the likelihood of the data given a model, and restrict our analyses to comparing the **relative** support for two or more models.
- The symmetry of the two metrics of support falls apart if $P(model|data)$ includes an *informative* prior probability distribution, rather than the constant *1/j*.

Graphical / geometric representation

Good video here: https://youtu.be/HZGCoVF3YvM