Lecture 5: Bayes pt. 1

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2016

Photo by Howard Jackman

Bayes

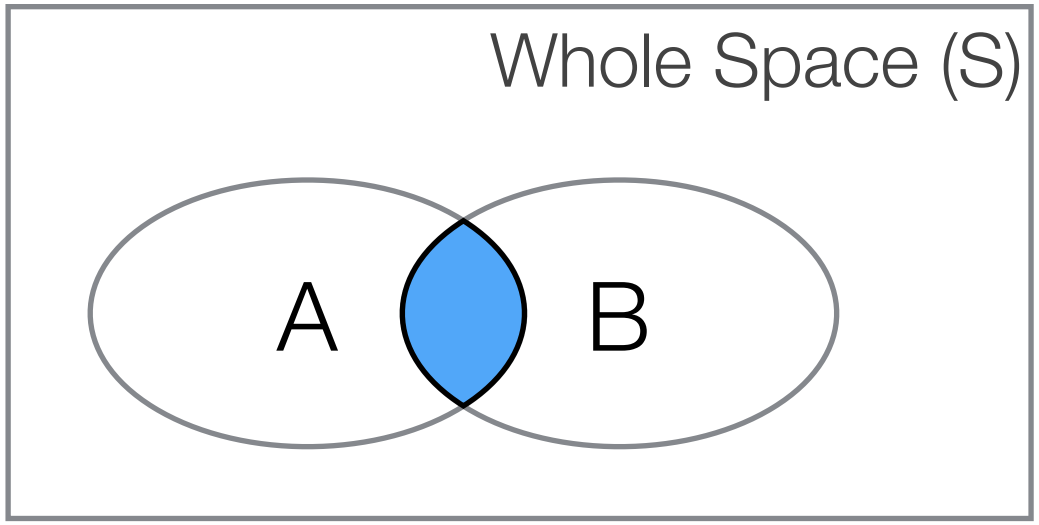
- Probabilities and statistics can encode an amount of belief in (data, model, systematics, hypothesis, parameters, etc.)
- Set notation
    - Not particularly nice but it has to be done
    - Comes with examples

Simple Notation

$A \cap B$ A intersect B

$B \cap A$ B intersect A

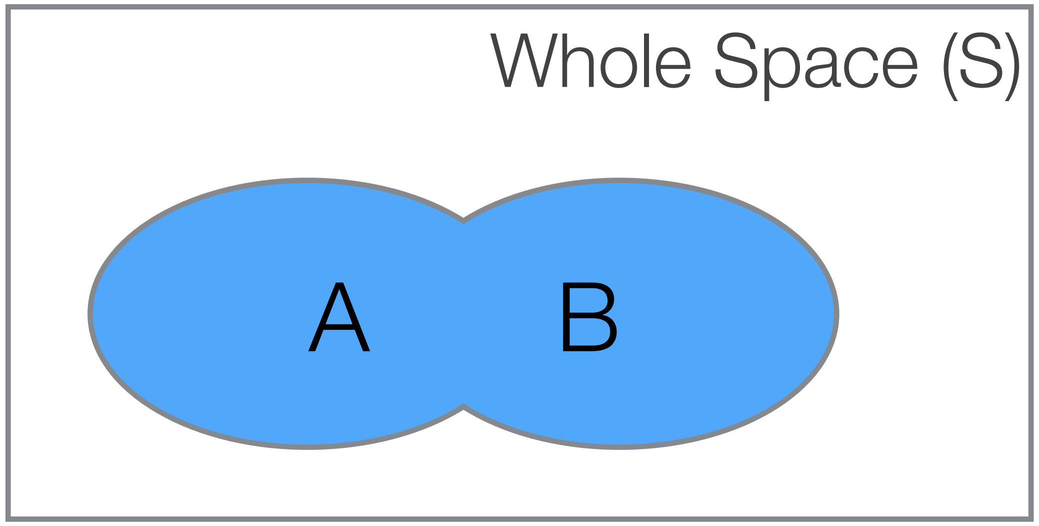
Only the things in both A and B



$$A \cap B = B \cap A$$

$A \cup B$ A union B

All the things in both A and B

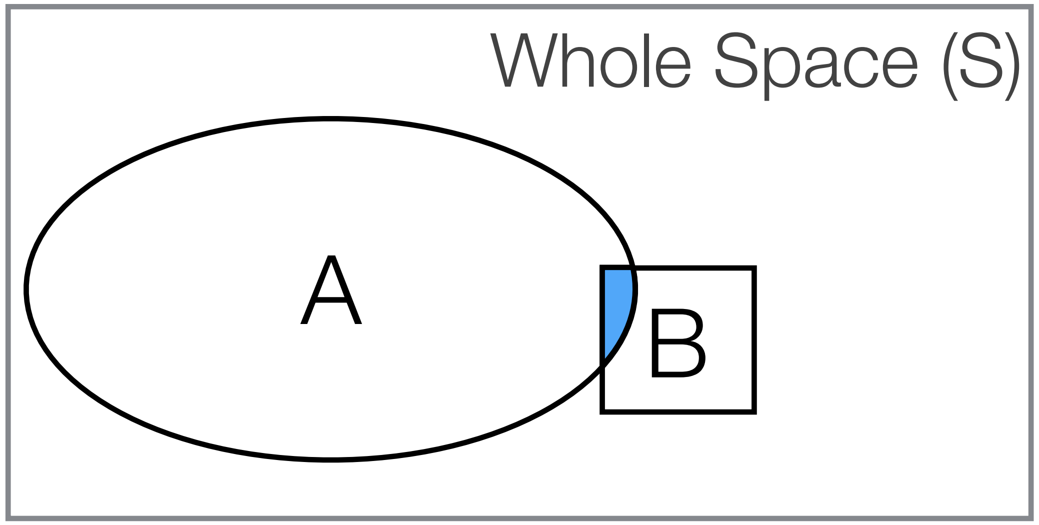


$$A \cup B = B \cup A$$

Moving to Probabilities

$P(A|B)$

"P of A given B"
or
"Probability of getting A given B"



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

- Intuitive, the probability of a parameter being in space A, given that the parameter is in space B, is the probability of the overlapping (intersect) space of A and B divided by the probability of being in space B
  - If intersect is large, so is the probability
  - If space B is small, and knowing that you are in space B, then the probability of being in A is large

Moving to Probabilities

- Rearranging some things :

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

$$A \cap B = B \cap A$$

$$P(A \cap B) = P(B \cap A) = P(A|B)P(B) = P(B|A)P(A)$$

- We get Bayes' theorem

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

- or sometimes

(Discrete)

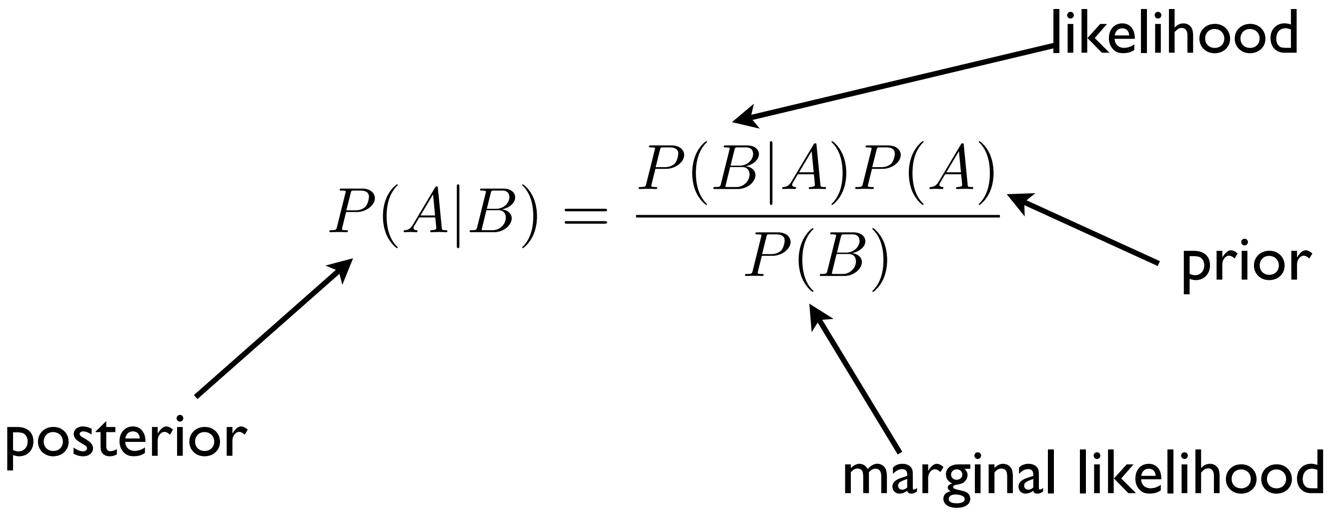
$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

(Continuous)

$$\frac{P(B|A)P(A)}{\int P(B|A)P(A)dA}$$

Bayes' Theorem

- One can solve the respective conditional probability equations for P(A and B) and P(B and A), setting them equal to give Bayes' theorem:



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

likelihood

prior

posterior

marginal likelihood

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

- The theorem applies to both frequentist and Bayesian methods. Differences stem from how the theorem is applied and, in particular, whether one extends probability to include some degree of belief.

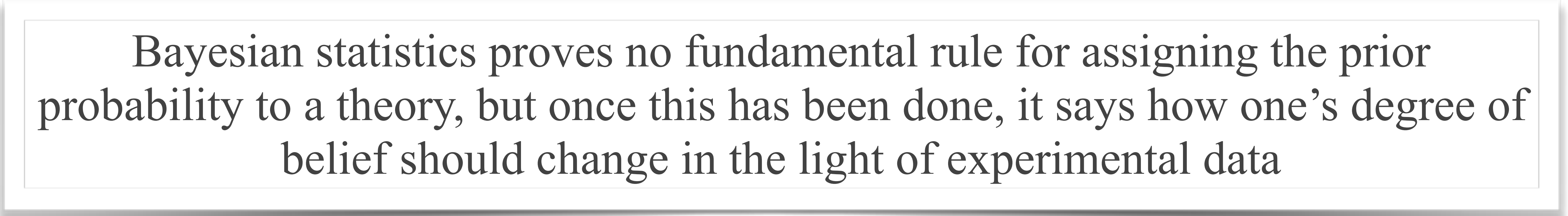
Interpretations

*T. Petersen

- One way Bayes' Theorem is often used in normal thinking is:

$$P(theory|data) \propto P(data|theory) \cdot P(theory)$$

- Here, P(data) has been omitted (doesn't depend on parameters. It contributes as a constant normalization and will be discussed later).
- The trouble being that it is hard to define P(theory) = a "degree of belief" in a theory.

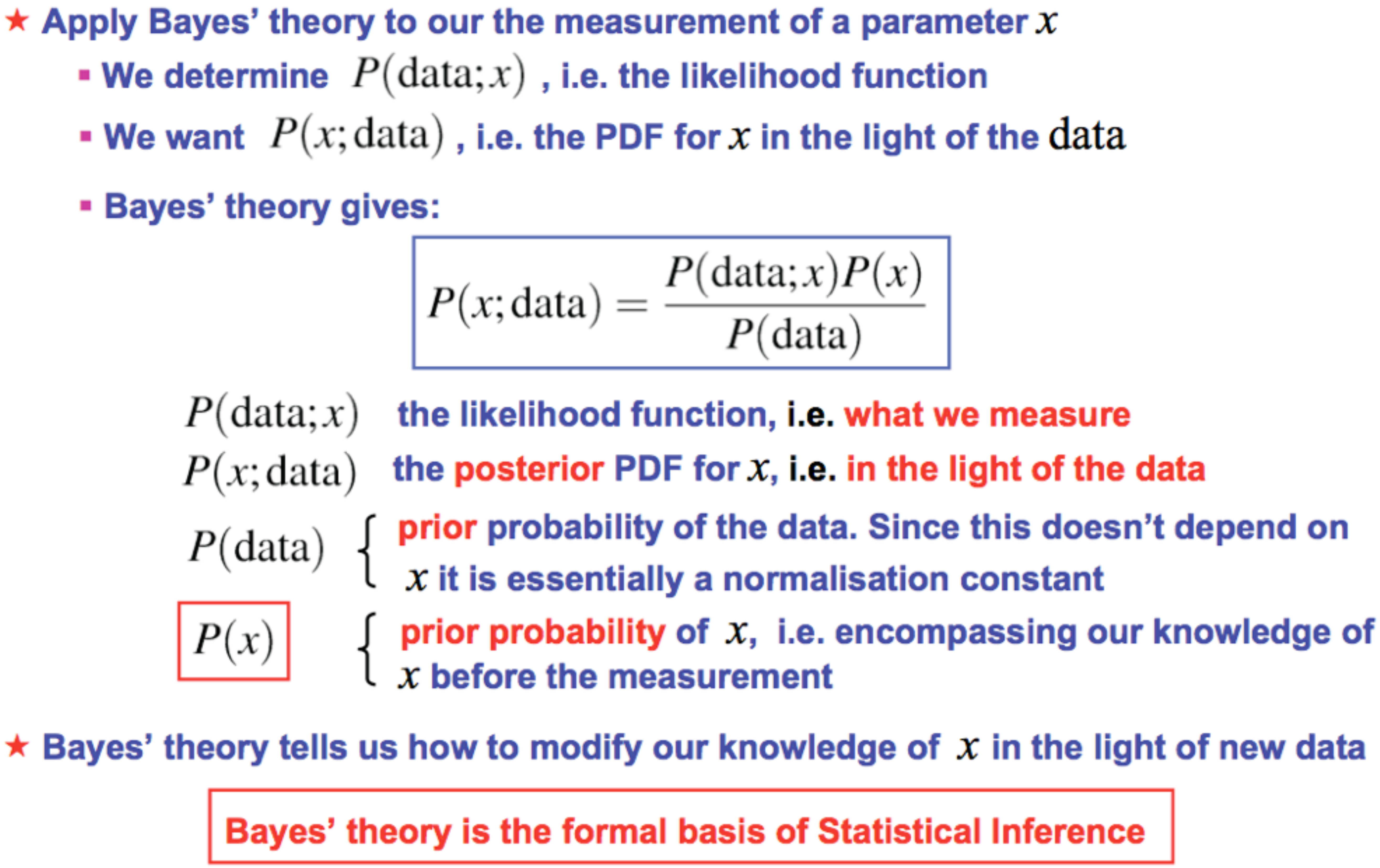


Bayesian statistics proves no fundamental rule for assigning the prior probability to a theory, but once this has been done, it says how one's degree of belief should change in the light of experimental data

-G. Cowan, "Statistical Data Analysis"

Application Overview

*M.A. Thomson



★ Apply Bayes' theory to our the measurement of a parameter x

- We determine $P(\text{data};x)$, i.e. the likelihood function
- We want $P(x;\text{data})$, i.e. the PDF for x in the light of the data
- Bayes' theory gives:

$$\boxed{P(x;\text{data}) = \frac{P(\text{data};x)P(x)}{P(\text{data})}}$$

$P(\text{data};x)$ the likelihood function, i.e. what we measure

$P(x;\text{data})$ the posterior PDF for x, i.e. in the light of the data

$P(\text{data})$ $\{$ prior probability of the data. Since this doesn't depend on x it is essentially a normalisation constant

$\boxed{P(x)}$ $\{$ prior probability of x, i.e. encompassing our knowledge of x before the measurement

★ Bayes' theory tells us how to modify our knowledge of x in the light of new data

Bayes' theory is the formal basis of Statistical Inference

Paradigm

- Inherently we are studying things that are unknown and how do you appropriately quantify the level of belief?
    - What is the prior on the speed of light in a vacuum being constant in all reference frames in 1900 versus 2016?
    - What is the amount of dark energy in the universe?
        - Prior on dark energy in the prevailing ΛCDM cosmological model?
        - Prior on ΛCDM?
- Bayesian (inference) statistics is not universally reasonable and applicable, nor is ‘frequentist’ statistics universally reasonable and applicable.

*Ex. 3.1 from "Modern Statistical Methods in Astronomy"

Astro Example

- In looking for radio loud (θ_1) versus quiet (θ_2) Active Galactic Nuclei (AGN) in a new patch of sky there is a devised test with the following likelihoods and priors:
  - The likelihood of correctly identifying a radio loud AGN is $P(+|\theta_1)=0.8$ while the likelihood of misidentifying (false positive) a non-radio loud AGN is $P(+|\theta_2)=0.3$
  - From previous studies, it is expected that the selected AGN in a new sky patch has a 10% radio loud AGN population, i.e. $P(\theta_1)=0.1$, and thus has 90% non-loud population $P(\theta_2)=0.9$
- What is the probability that an AGN that is observed as radio loud is actually radio loud, $P(\theta_1|+)$?
- What's better, decreasing the false positives by a factor of 2 or improving the AGN selection by 60%?

*Ex. 3.1 from “Modern Statistical Methods in Astronomy”

Answers

- What is the probability that an AGN that is observed as radio loud is actually radio loud, P(θ_1|+)?

$$P(\theta_1|+) = \frac{0.8 \times 0.1}{0.8 \times 0.1 + 0.3 \times 0.9} \approx 0.229$$

- What’s better, decreasing the false positives by a factor of 2 or improving the AGN selection by 60%?

$$P(\theta_1|+) = \frac{0.8 \times (0.16)}{0.8 \times (1.6) + 0.3 \times (0.84)} \approx 0.337$$

$$P(\theta_1|+) = \frac{0.8 \times 0.1}{0.8 \times 0.1 + (0.3/2) \times 0.9} \approx 0.372$$

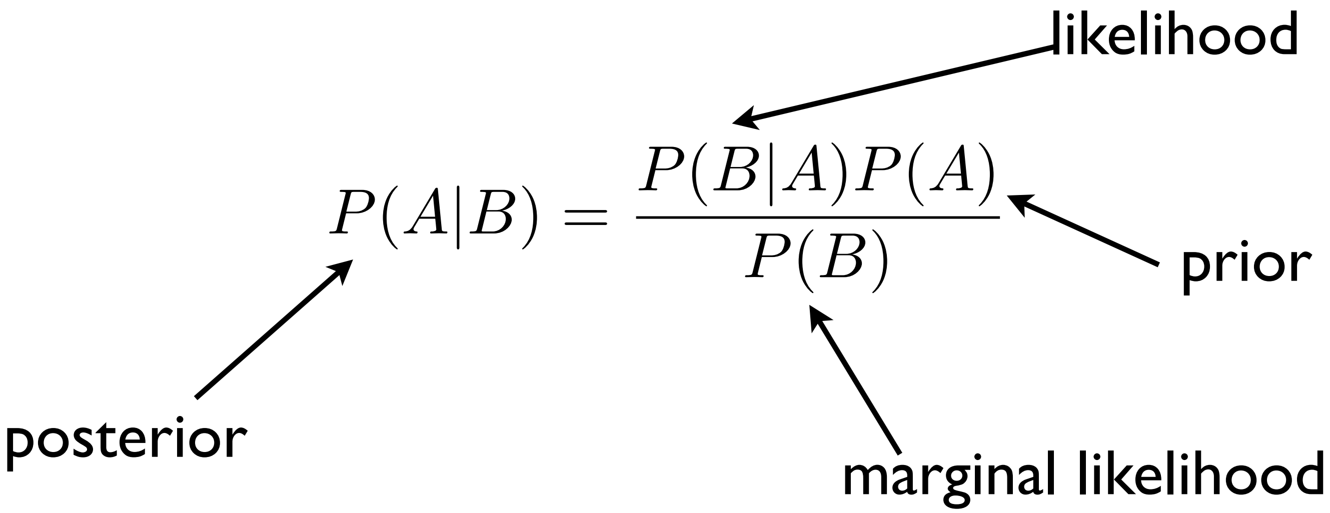
Reducing false positives by factor 2 is better

Exercise #1

- We want to find out the population of N identical things, e.g. fish, cancer cells, gas in a semi-evacuated volume, etc. We extract *n* that are identified (tagged, radioactive marked, isotope altered, etc.) and released back into the population. After sufficiently re-mixing long, *K* things are extracted and checked as to whether they have been previously tagged (*k*).
  - For *n*=60, *K*=100, and *k*=10, what is the total population (N)?
  - Natural guess is N=100/10*60, but that gives only a single number. What we want is the posterior distribution which provides more information.

Exercise #1 (cont.)

- P(N|k) is the posterior and gives us the conditional probability of the total population N, given k=10.
- Using Bayes' theorem, and knowing that we have data where k=10, K=100, and n=60 the posterior is proportional to the likelihood of P(k|N)
    - P(k|N) is a 'sampling w/o replacement' likelihood and is a hypergeometric probability (go online and find the quasi-binomial form of this likelihood)
    - We will come back to the marginal likelihood later, but for here pick a fixed number.
    - The posterior is then produced by using P(k|N) and scanning across values of N.
    - Using a flat prior, i.e. prior is constant, plot the posterior distribution



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

likelihood

prior

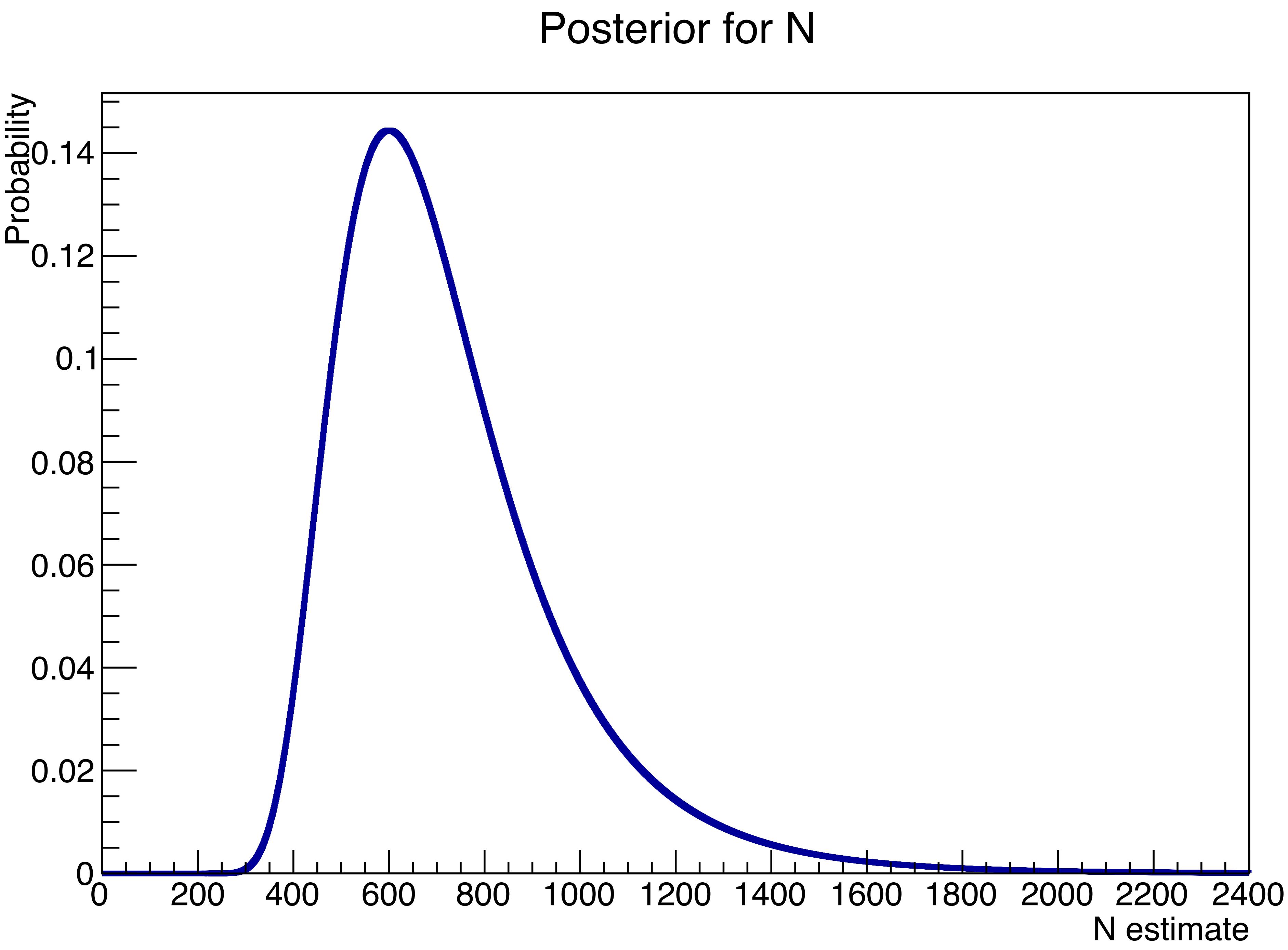
posterior

marginal likelihood

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Exercise #1 (cont.)

- Here is the posterior distribution for a flat prior of 2.



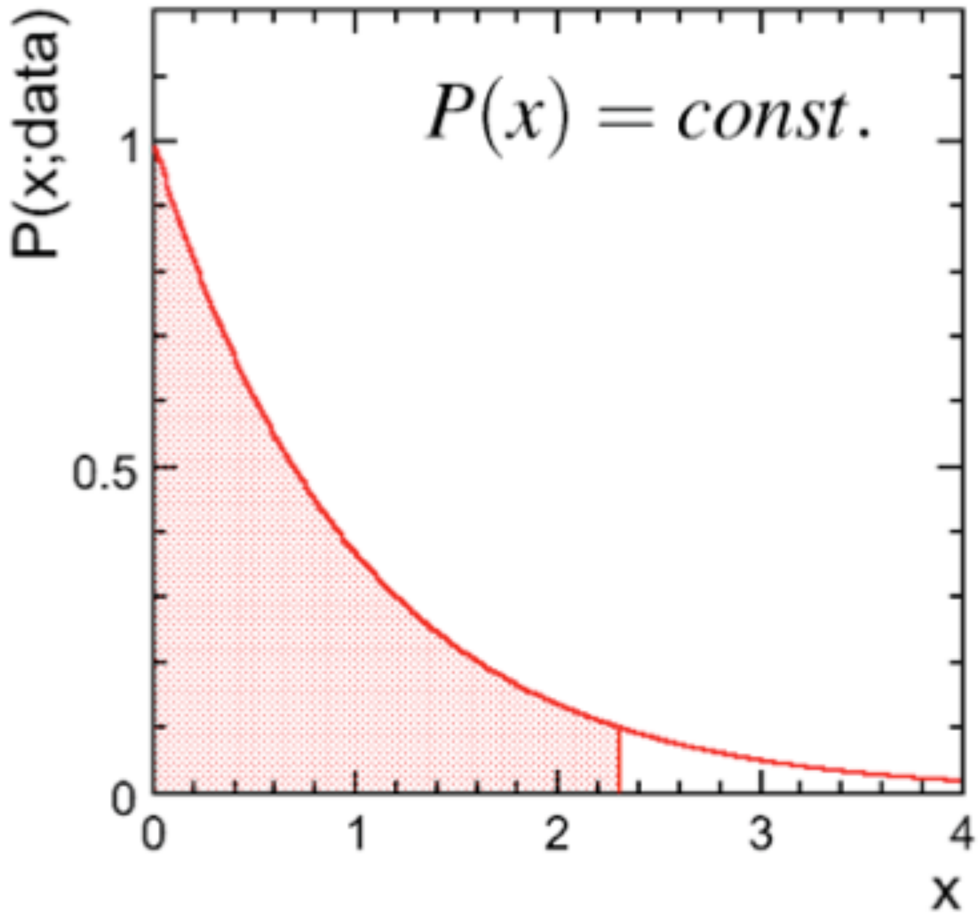
Importance of Priors

- ★ **See no events…**

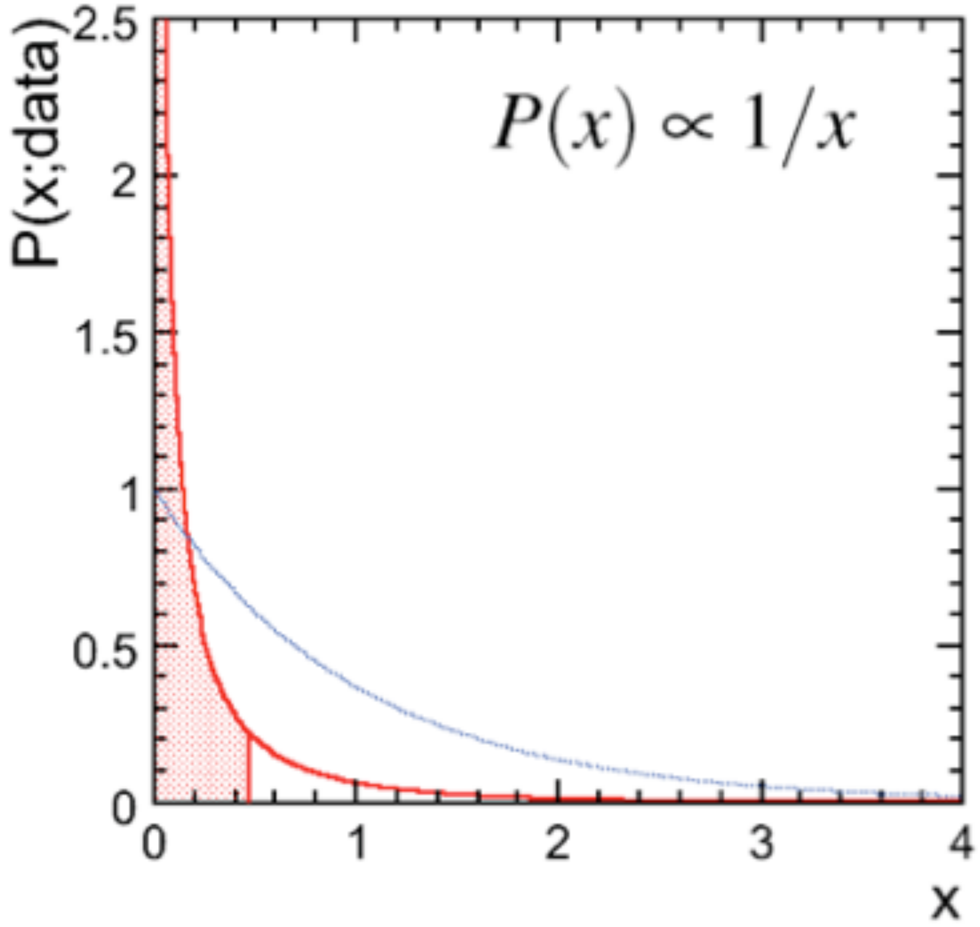
$$P(\text{data};x) = P(0;x) = e^{-x}$$

Poisson prob. for observing 0

Prior flat prior in x : $P(x) = const.$



Prior flat prior in lnx : $P(\ln x) = const.$



- ★ **The Conclusions are very different. Compare regions containing 90 % of probability**

$x < 2.3$ $x < 0.46$

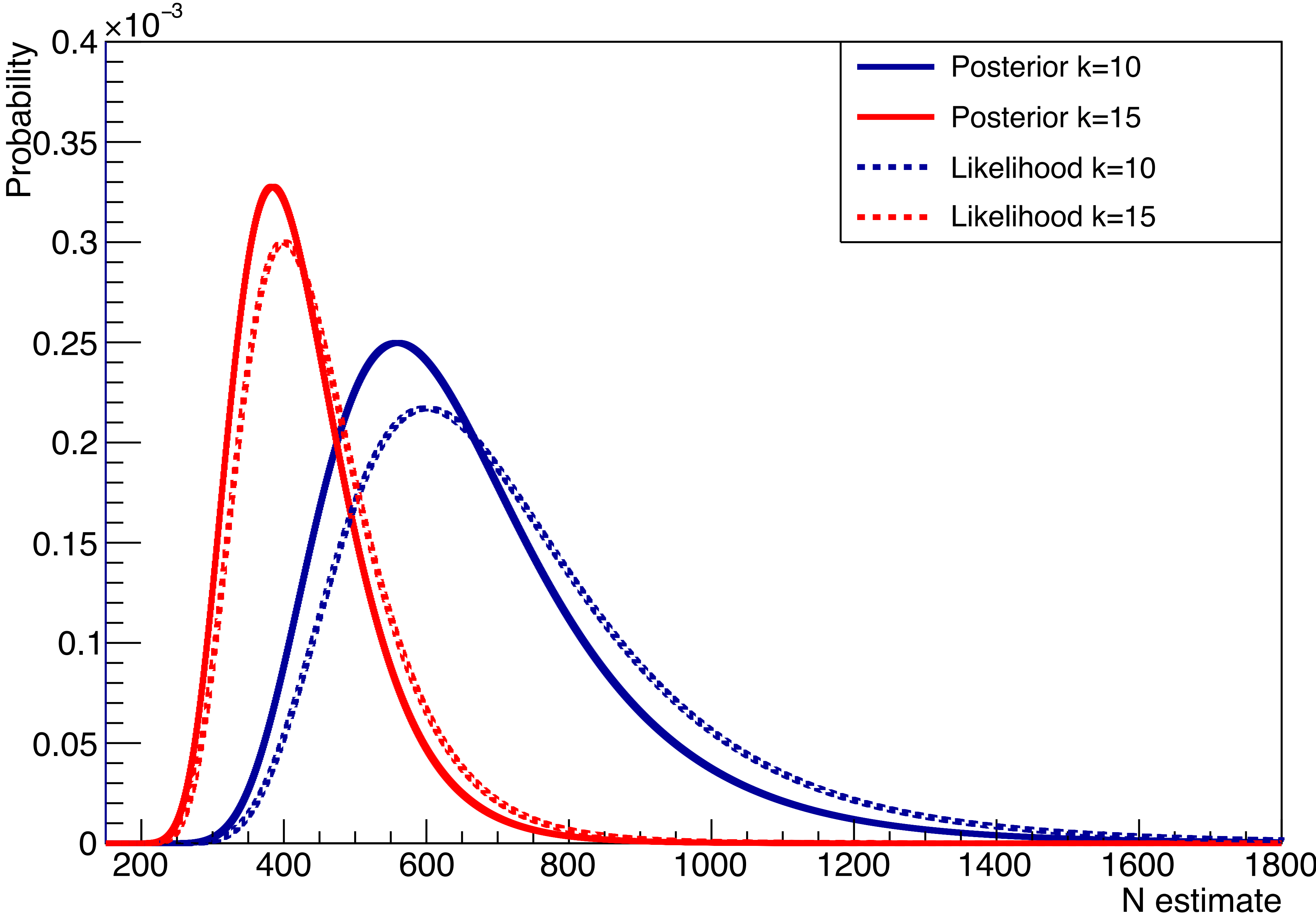
- **In this case, the choice of prior is important**

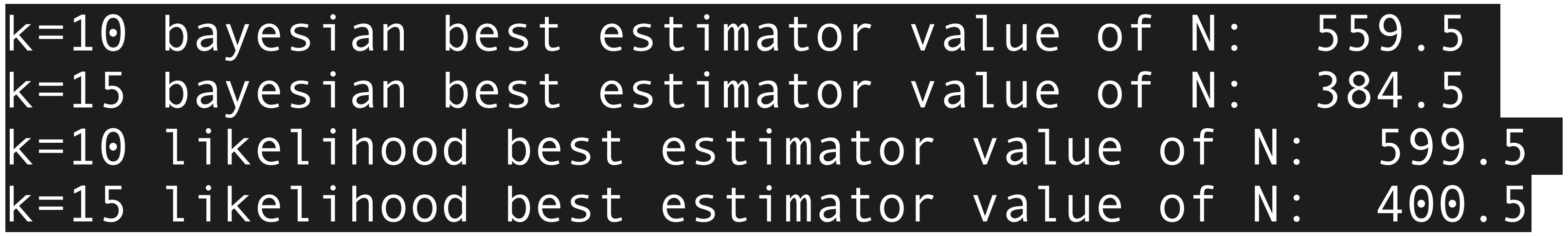
Exercise #2

- With the posterior distribution for k=10, also plot for k=15
- Because the posteriors are proportional to the likelihoods*prior, also plot the likelihoods on the same plots as the 2 posterior distributions
    - For a flat prior
    - For a prior of the form 1/N
    - Don’t worry too much right now about normalization of priors, likelihoods, and/or posteriors, just make sure they show up on a ‘reasonable’ scale for the plots
- Do the estimator values for N differ between the likelihoods and the Bayesian posteriors for a flat prior?
    - What about for the 1/N prior?

Exercise #2 (cont.)

- Prior is 1/N
- Values in terms of 0.5 are a binning/ histogram artifact, so don't be perplexed if your values are slightly different





```
k=10 bayesian best estimator value of N:   559.5
k=15 bayesian best estimator value of N:   384.5
k=10 likelihood best estimator value of N:   599.5
k=15 likelihood best estimator value of N:   400.5
```

Exercise Fish (cont.)

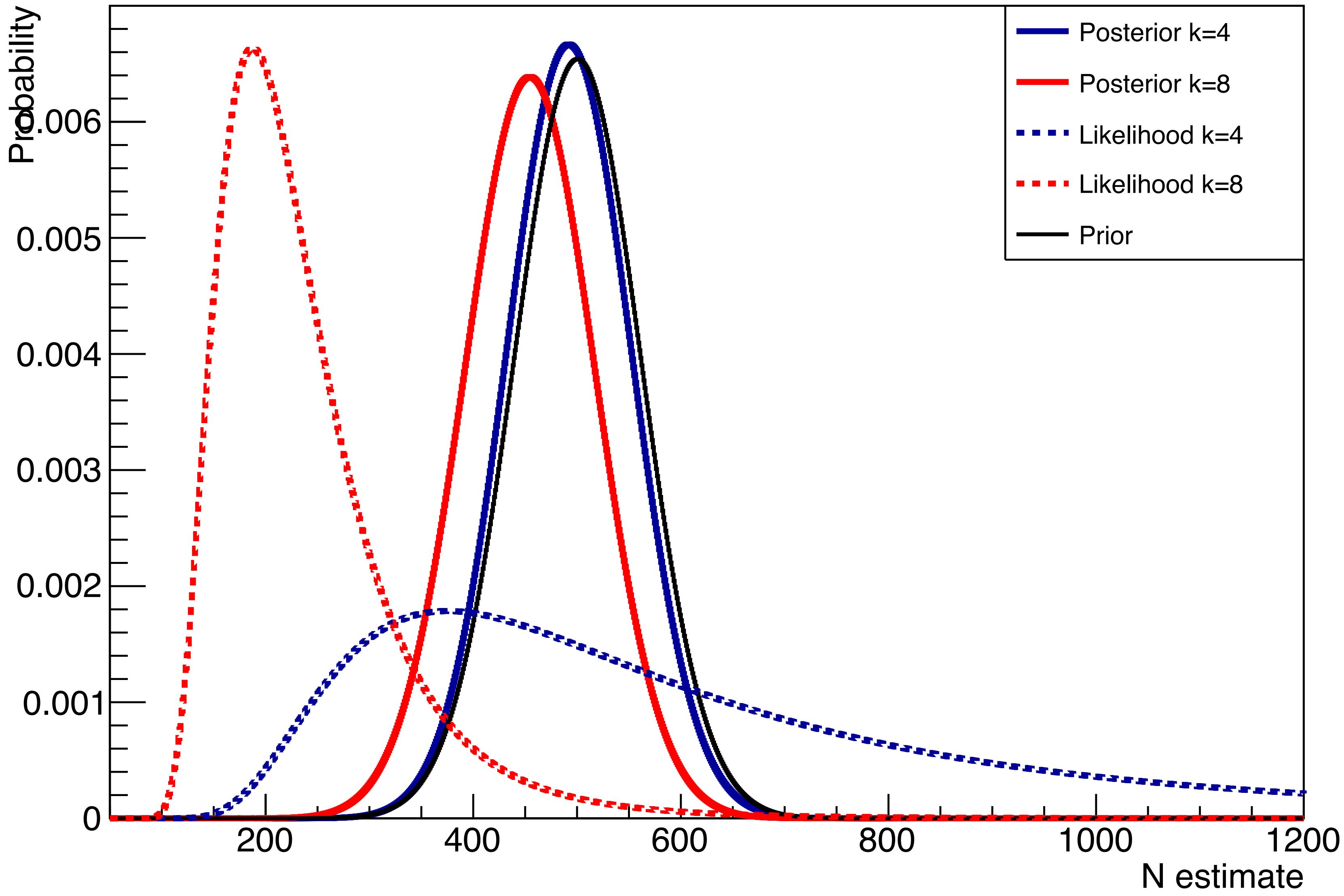
- The previous priors were not necessarily too well informed, so let's use a less abstract scenario where we can include a more informed prior
- Estimate the population (N) of a species of fish in a lake. Assume that between tagging of n fish and re-sampling there is enough time for sufficient mixing, but not enough to alter the total population, e.g. reduction from predators, births, pollution induced die-off, etc.
  - From other studies you know that the fish prefers 10±1 m^3 of water free from any fish of the similar species and that the entire lake is 5000±300 m^3
  - You hypothesize that the population has saturated, so there are ~500 fish in the lake

Exercise Fish (cont.)

- With a hypothesized mean of fish, form a gaussian prior based on the uncertainties related to the volume approximation of both the entire lake as well as the volume that the fish prefer (for simplicity we'll assume they're aggressive fish and, except for mating, kill any other similar species fish which come within their 10 m^3 volume). For a K=50, n=30, and k=4:
  - How sensitive is the posterior to the form or values in the prior?
  - Repeat with only changing from k=4 to k=8.
  - If the gaussian uncertainty is doubled or tripled, how much closer are the likelihood estimator values to the posterior estimator values for k=4 and k=8?
  - What if another study suggests that the fish actually prefer 9.2±0.2 m^3

Posterior/Likelihood Differences

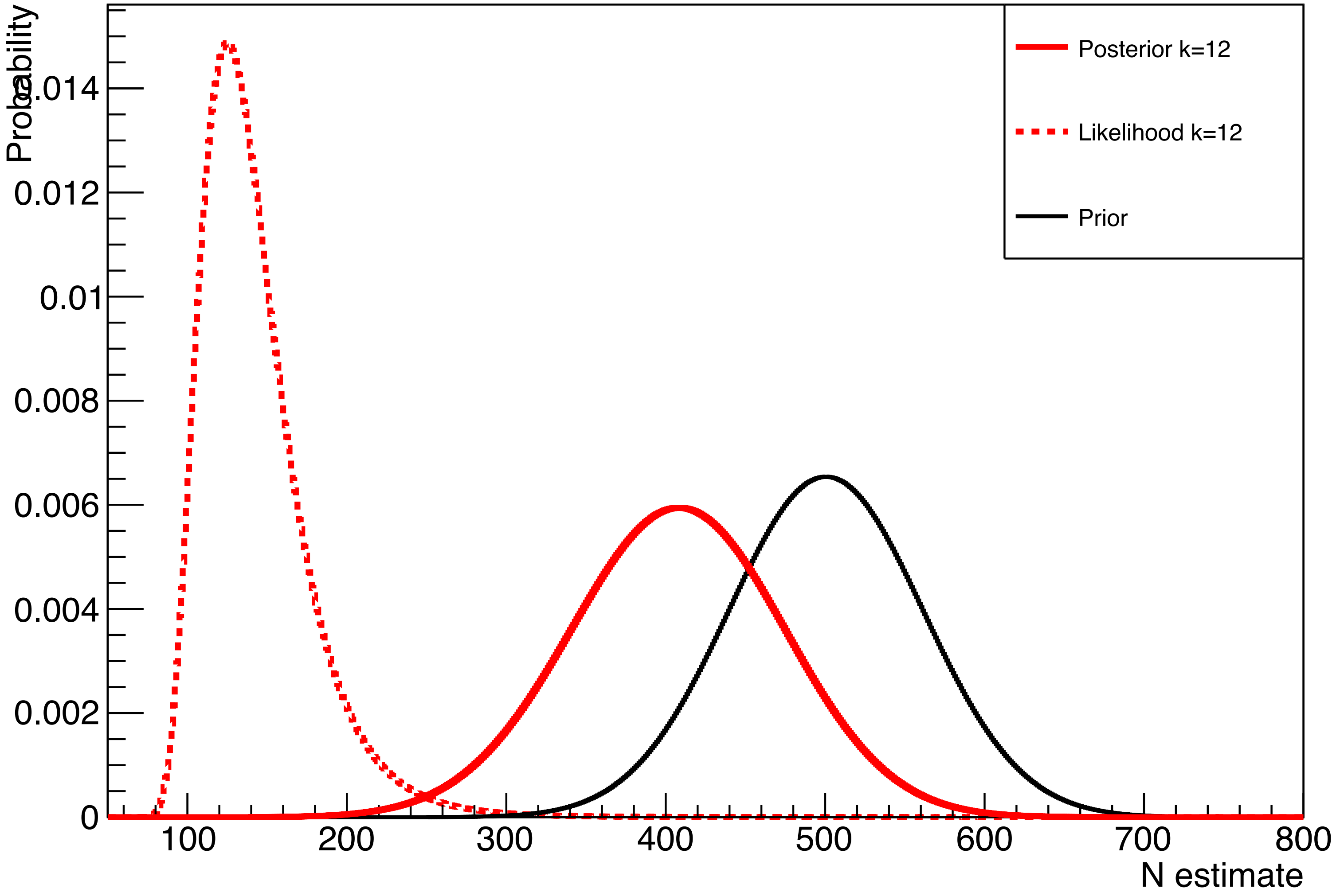
- For the instance where k=8, gaussian mean=500 and σ=61, we can see a large separation between the posterior and likelihood



Even More Extreme

- For the instance where k=12, gaussian mean=500 and σ=61 we've got some some issues

- The bayesian posterior best estimate is ~409, but the best likelihood estimate is ~125.
- According to the likelihood PDF, how likely is it to have a value ≥ 409?
  - (hint integrate the tail of the likelihood distribution ≥ 409)



Answers

- According to my code and the likelihood PDF, there is a probability of ~0.00017 of randomly getting a value ≥409, which is the most likely value according to the bayesian posterior.

Answers

- According to my code and the likelihood PDF, there is a probability of ~0.00017 of randomly getting a value ≥409, which is the most likely value according to the bayesian posterior.

No, not exactly.

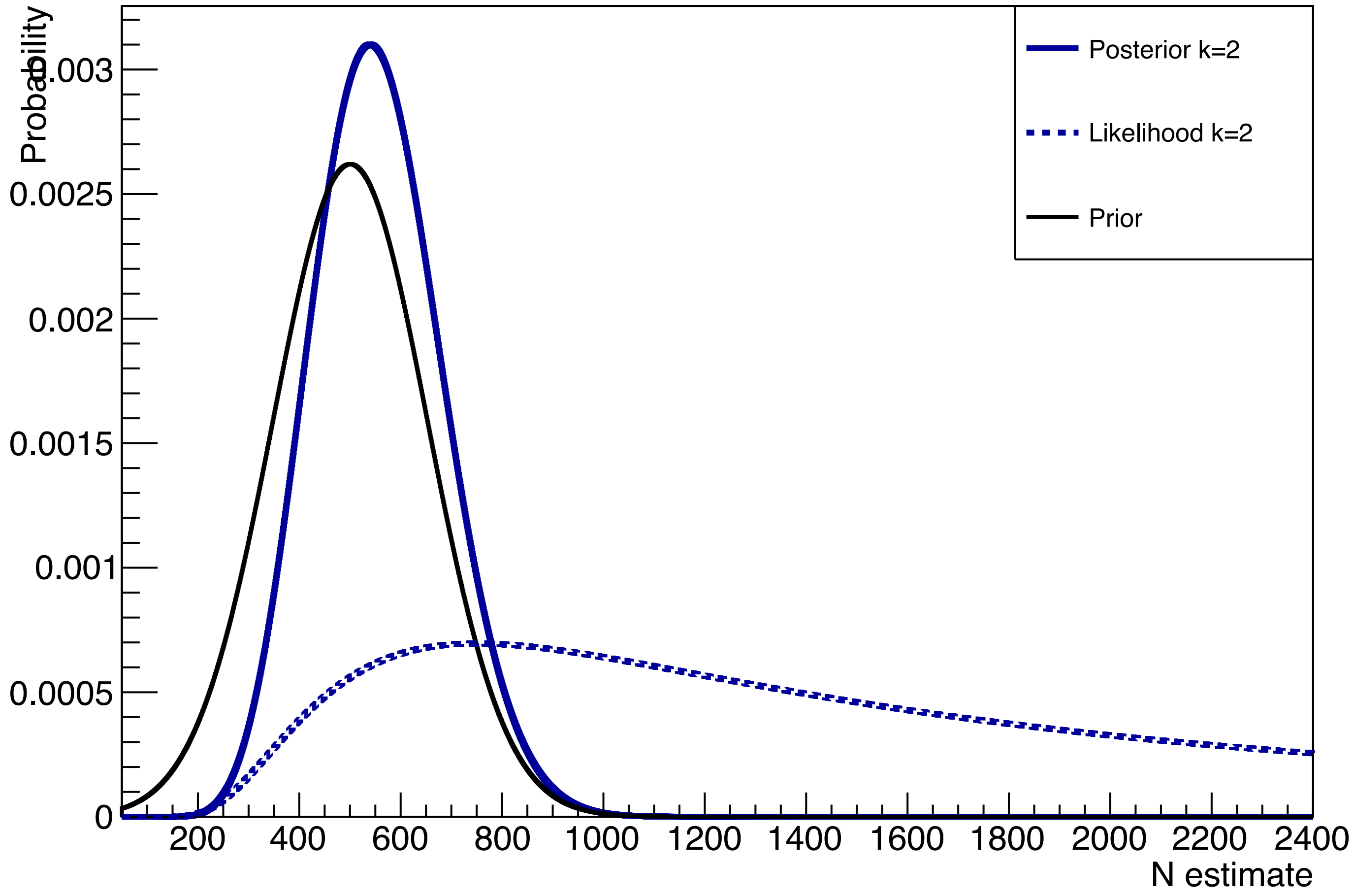
*Dumb and Dumber

Answers

- According to my code and the likelihood PDF, there is a probability of ~0.00017 of randomly getting a value ≥409, which is the most likely value according to the bayesian posterior.
- With such a divergence between the likelihood, posterior, and dependence on the prior, it is worth investigating whether the prior (or it's parameters) or the likelihood are appropriate.
- The most likely bayesian estimator, i.e. the mode, may not be the best test metric versus using the median or mean. But it's clear from the plot that any metric comparing the posterior to the likelihood should not find great compatibility.

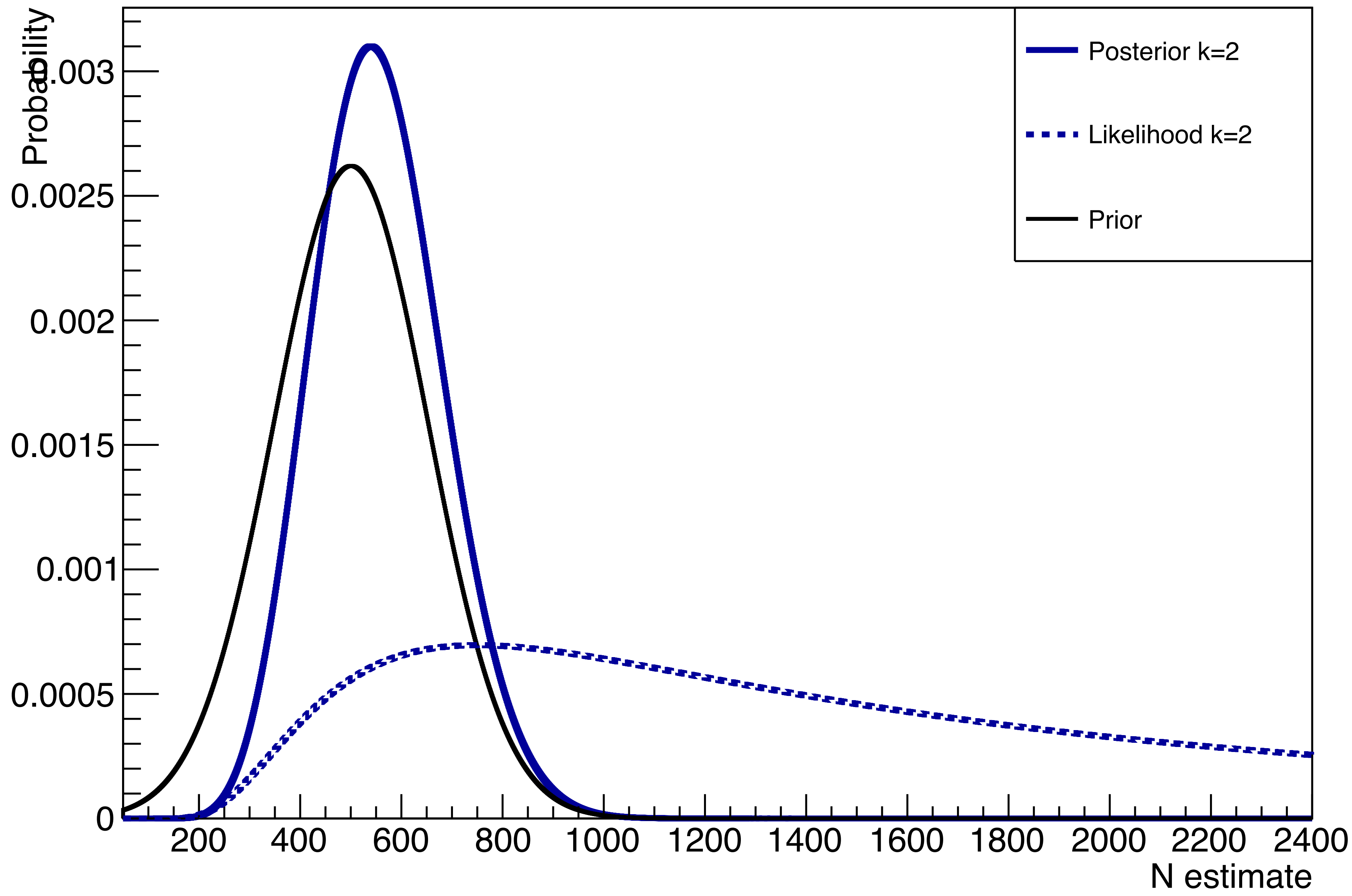
Other Extrema

- For the instance where k=2, gaussian mean=500 and σ=152.5 we have a likelihood function that has a very wide range of near equally likely fish populations. This not terribly informative.
- By including prior information it is possible to encode some belief that can provide useful information. But, the result is sensitive to the prior, because the data is not very discriminating.



Other Extrema

- If the prior is very well justified, then the data dramatically improves the population estimates < ~250.



Additional

- The form of the previous priors were all gaussian, but there are many, many more options
  - For a larger list http://www.fysik.su.se/~walck/suf9601.pdf
  - What happens when you switch to a beta-distribution of similar shape as the prior? (Here you can eyeball and tune so that your gaussian and beta-distribution are ‘similar’)
- At some point the lake is full of fish, or at least beyond realistically populated. What happens to the posterior with a truncated gaussian prior?
  - Truncated at 50,000 fish?
  - Truncated at 2,000 total fish?
- You’ll notice that there is a lower limit on the number of fish related to some combination of K, k, and n, which was not included in the previous exercise results. What do they look like if the lower bound is included?

Markov Chain Monte Carlo

- Next Class