Choosing Priors
Probability Intervals
18.05 Spring 2022

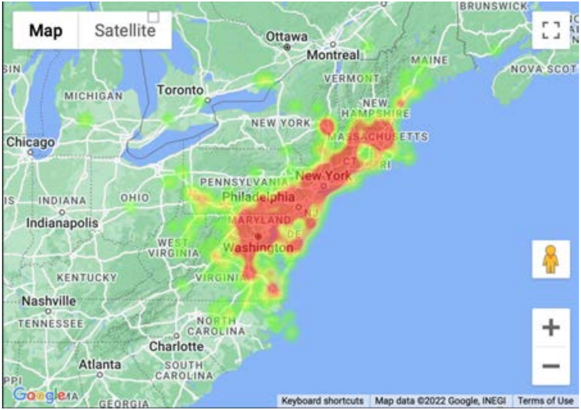


Image created from data taken from https://amsmeteors.org/2013/03/update-for-march-22-2013-northeast-fireball

Announcements/Agenda

Announcements

- None

Agenda

- Two parameter tables.
- Choosing priors
- Probably only one board question today
- Left meteor slides in at the end for fun. We probably won't have time to cover them.

Two-parameter tables: Malaria

In the 1950s scientists injected 30 African "volunteers" with malaria.

S = carrier of sickle-cell gene

N = non-carrier of sickle-cell gene

$D+$ = developed malaria

$D-$ = did not develop malaria

	$D+$	$D-$	
S	2	13	15
N	14	1	15
	16	14	30

Model

θ_S = probability an injected S develops malaria.

θ_N = probability an injected N develops malaria.

Assume conditional independence between all the experimental subjects.

Likelihood is a function of both θ_S and θ_N:

$$P(\text{data}|\theta_S, \theta_N) = c\,\theta_S^2(1-\theta_S)^{13}\theta_N^{14}(1-\theta_N).$$

Hypotheses: pairs (θ_S, θ_N).

To simplify, use a finite number of hypotheses:
θ_S and θ_N are each one of $0, 0.2, 0.4, 0.6, 0.8, 1$.

Color-coded two-dimensional tables

Hypotheses

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1
1	(0,1)	(.2,1)	(.4,1)	(.6,1)	(.8,1)	(1,1)
0.8	(0,.8)	(.2,.8)	(.4,.8)	(.6,.8)	(.8,.8)	(1,.8)
0.6	(0,.6)	(.2,.6)	(.4,.6)	(.6,.6)	(.8,.6)	(1,.6)
0.4	(0,.4)	(.2,.4)	(.4,.4)	(.6,.4)	(.8,.4)	(1,.4)
0.2	(0,.2)	(.2,.2)	(.4,.2)	(.6,.2)	(.8,.2)	(1,.2)
0	(0,0)	(.2,0)	(.4,0)	(.6,0)	(.8,0)	(1,0)

Table of hypotheses for (θ_S, θ_N)

Corresponding level of protection due to S: $\theta_N - \theta_S$.

orange = strong, darker blue = some, light blue = none, white = negative.

Color-coded two-dimensional tables

Likelihoods (scaled to make the table readable)

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	1.93428	0.18381	0.00213	0.00000	0.00000
0.6	0.00000	0.06893	0.00655	0.00008	0.00000	0.00000
0.4	0.00000	0.00035	0.00003	0.00000	0.00000	0.00000
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Likelihoods scaled by $100000/c$

$$p(\text{data}|\theta_S, \theta_N) = c\,\theta_S^2(1-\theta_S)^{13}\theta_N^{14}(1-\theta_N).$$

High likelihood hypotheses inside thick black rectangle.

Color-coded two-dimensional tables

Flat prior

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.8	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(\theta_S)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Flat prior $p(\theta_S, \theta_N)$: each hypothesis (square) has equal probability

Color-coded two-dimensional tables

Posterior to the flat prior

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N \mid \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88075	0.08370	0.00097	0.00000	0.00000	0.96542
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00002	0.00000	0.00000	0.00000	0.00018
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S \mid \text{data})$	0.00000	0.91230	0.08670	0.00100	0.00000	0.00000	1.00000

Normalized posterior to the flat prior: $p(\theta_S, \theta_N \mid \text{data})$

Strong protection:
$P(\theta_N - \theta_S \geq 0.6 \mid \text{data}) = \text{sum of orange} = 0.88075$

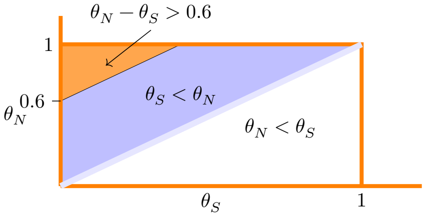
Some protection:
$P(\theta_N > \theta_S \mid \text{data}) = \text{sum orange and darker blue} = 0.99995$

Continuous two-parameter distributions

(This is preparation for the board problem.)
Sometimes continuous parameters are more natural.

Malaria example (from class notes):
discrete prior table from the class notes.
Similarly colored version for the continuous parameters (θ_S, θ_N) over range $[0,1] \times [0,1]$.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1
1	(0,1)	(.2,1)	(.4,1)	(.6,1)	(.8,1)	(1,1)
0.8	(0,.8)	(.2,.8)	(.4,.8)	(.6,.8)	(.8,.8)	(1,.8)
0.6	(0,.6)	(.2,.6)	(.4,.6)	(.6,.6)	(.8,.6)	(1,.6)
0.4	(0,.4)	(.2,.4)	(.4,.4)	(.6,.4)	(.8,.4)	(1,.4)
0.2	(0,.2)	(.2,.2)	(.4,.2)	(.6,.2)	(.8,.2)	(1,.2)
0	(0,0)	(.2,0)	(.4,0)	(.6,0)	(.8,0)	(1,0)



The probabilities are given by double integrals over regions.

Board question preamble: Treating severe respiratory failure*

Two treatments for newborns with severe respiratory failure.

1. CVT: conventional therapy (hyperventilation and drugs)

2. ECMO: extracorporeal membrane oxygenation (invasive procedure)

In 1983 in Michigan:

19/19 ECMO babies survived and 0/3 CVT babies survived.

Later Harvard ran a randomized study:

28/29 ECMO babies survived and 6/10 CVT babies survived.

*Adapted from *Statistics: a Bayesian Perspective* by Donald Berry

Board question: Updating two parameter priors

Michigan: 19/19 ECMO babies and 0/3 CVT babies survived.

Harvard: 28/29 ECMO babies and 6/10 CVT babies survived.

θ_E = probability that an ECMO baby survives
θ_C = probability that a CVT baby survives.

Consider the values 0.125, 0.375, 0.625, 0.875 for θ_E and θ_C.

(a) Make the 4×4 prior table for a flat prior.
(b) Based on the Michigan results, create a reasonable informed prior table for analyzing the Harvard results (unnormalized is fine).
(c) Make the likelihood table for the Harvard results. (You might use R to compute some of the values.)
(d) Find the posterior table for the informed prior.
(e) Using the informed posterior, compute the probability that ECMO is better than CVT.
(f) Also compute the posterior probability that $\theta_E - \theta_C \geq 0.6$.
(The posted solutions will also show 4-6 for the flat prior.)

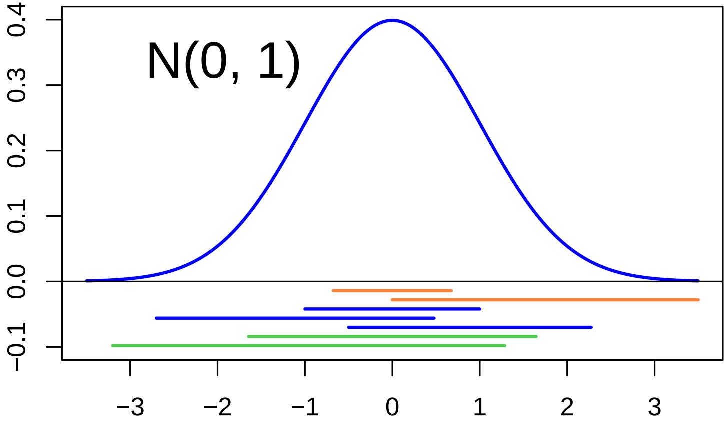
Probability intervals

- **Example.** If $P(a \leq \theta \leq b) = 0.7$ then $[a, b]$ is a 0.7 probability interval for θ. We also call it a 70% probability interval.
- **Example.** Between the 0.05 and 0.55 quantiles is a 0.5 probability interval. Another 50% probability interval goes from the 0.25 to the 0.75 quantiles.
- **Symmetric probability intervals.** A symmetric 90% probability interval goes from the 0.05 to the 0.95 quantile.
- **Q-notation.** Writing q_p for the p quantile we have 0.5 probability intervals $[q_{0.25}, q_{0.75}]$ and $[q_{0.05}, q_{0.55}]$.
- **Uses.** To summarize a distribution; to help build a subjective prior.

Probability intervals in Bayesian updating

- We have p-probability intervals for the prior $f(\theta)$.
- We have p-probability intervals for the posterior $f(\theta|x)$.
- The latter tend to be smaller than the former. Thanks data!
- Probability intervals are good, concise statements about our current belief/understanding of the parameter of interest.
- We can use them to help choose a good prior.

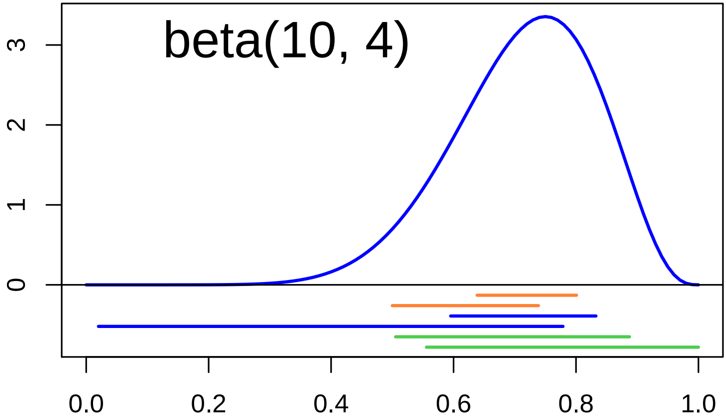
Probability intervals for normal distributions



orange = 0.50, blue = 0.68, green = 0.9

Note. The asymmetric intervals are of different lengths.

Probability intervals for beta distributions



orange = 0.50, blue = 0.68, green = 0.9

Concept question: Increasing probability

To convert an 80% probability interval to a 90% interval should you shrink it or stretch it?

1. Shrink **2.** Stretch.

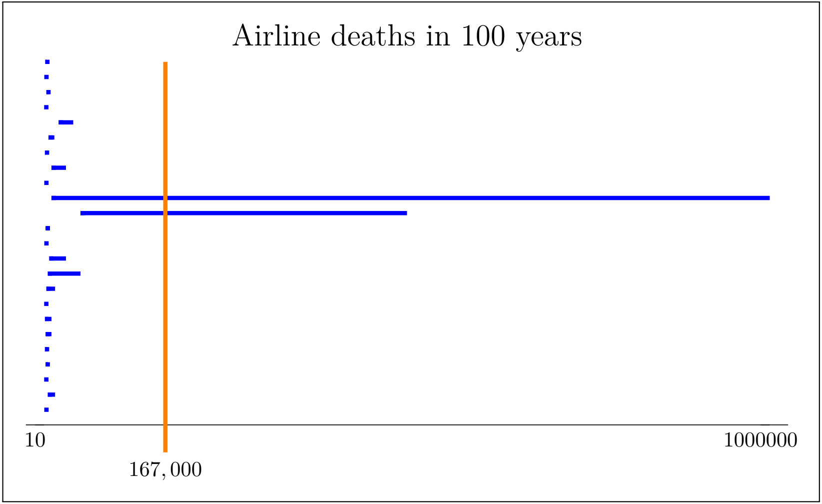
Reading questions

The following slides contain bar graphs of the responses to the reading questions. Each bar represents one student's estimate of their own 50% probability interval (from the 0.25 quantile to the 0.75 quantile).

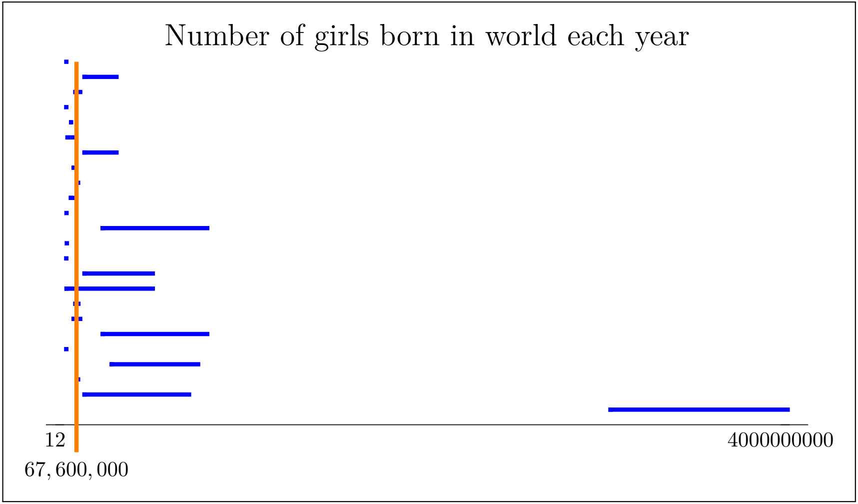
Each graph also shows the 'correct' answer gleaned from a search of the web.

Details about these answers are in the slides following the graphs.

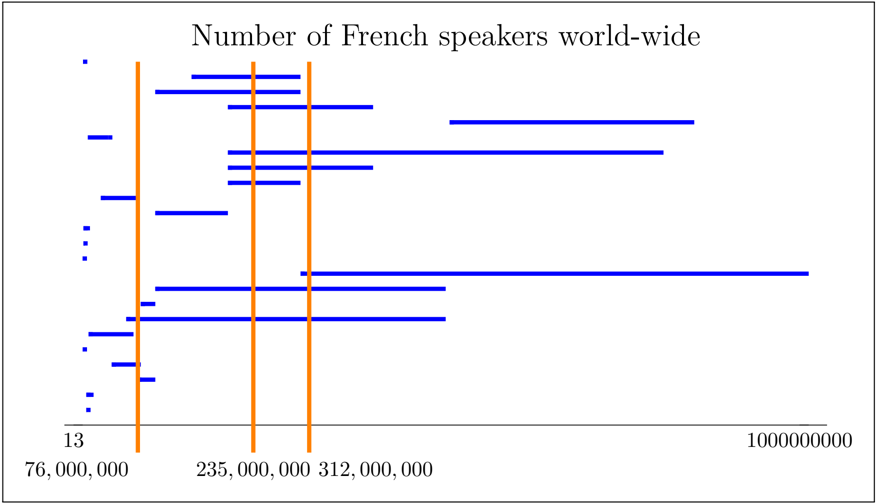
Subjective probability 1 (50% probability interval)



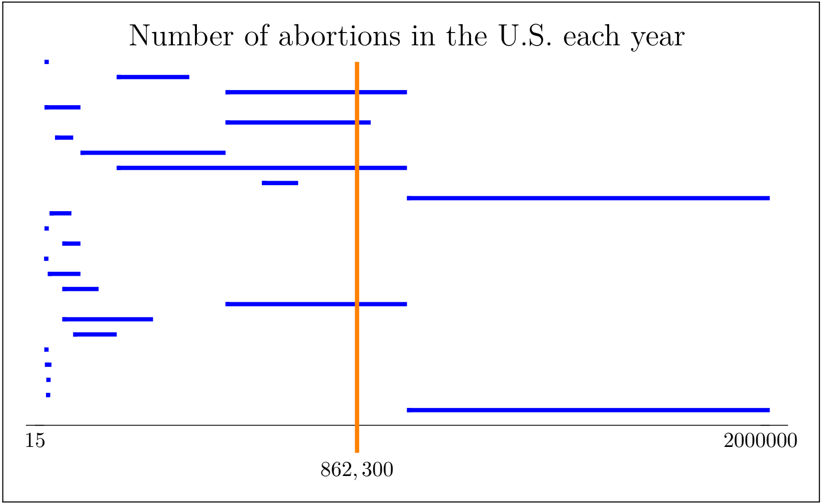
Subjective probability 2 (50% probability interval)



Subjective probability 3 (50% probability interval)



Subjective probability 4 (50% probability interval)



Reading question answers

Here is what we found for answers to the questions:

1. Airline deaths in 100 years: We found a Wikipedia article that claimed there were 83,772 fatalities due to aviation accidents between 1970 and 2020 (https://en.wikipedia.org/w/index.php?title=Aviation_accidents_and_incidents&oldid=1079179593).
In the 50 years before that there were fewer people flying, but it was probably more dangerous. Let's assume they balance out and estimate the total number of fatalities in 100 years as $2 \times 83,772 \approx 167,000$

2. Number of girls born in the world each year: I had trouble finding a totally reliable source, but most sites gave the same number of, currently, about 140 million births each year, e.g. https://ourworldindata.org/grapher/births-and-deaths-projected-to-2100. If we take what seems to be the accepted ratio of 1.07 boys born for every girl then $140/2.07 = 67.6$ million baby girls.

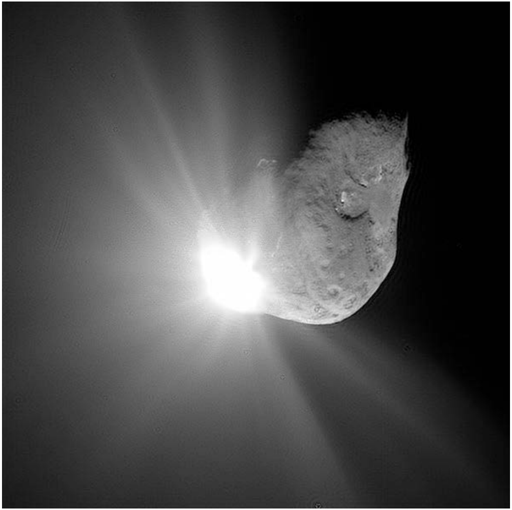
Reading question answers continued

3. It's hard to get precise numbers but all the sites I found gave similar numbers: For instance, https://en.wikipedia.org/w/index.php?title=French_language&oldid=1080625160 says there about 76 million native speakers, 235 million daily fluent speakers and and additional 77-110 million secondary speakers who speak it with varying degrees of proficiency.

4. According to Wikipedia (https://en.wikipedia.org/w/index.php?title=Abortion_statistics_in_the_United_States&oldid=1080417881), which is quoting Guttmacher Institute figures, there were 862,300 abortions in the US in 2017. They also quote the CDC figure of 619,591. The difference probably reflects the fact that the CDC is using official figures reported by the states and not all states report each year. Also, they are only reporting legal, induced abortions, while Guttmacher is using a more comprehensive survey.

Meteor!

On March 22, 2013, a meteor lit up the skies.
It passed almost directly over NYC.



Public domain image courtesy of NASA.

Board question: Meteor! (a) No data.

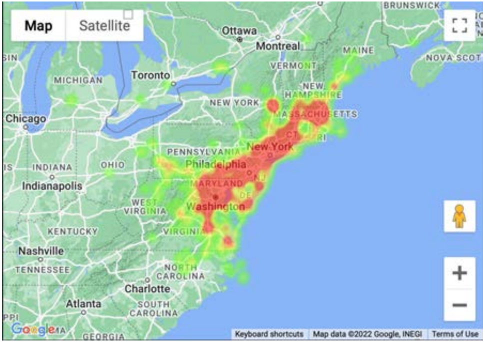
Draw a pdf $f(\theta)$, $0 \leq \theta < 2\pi$, for the meteor's direction.

Draw a 0.5-probability interval. How long is it?

Let $\theta = 0$ mean due east.

(Stay at the board for the next questions.)

Board question: Meteor! (b) Heat map.

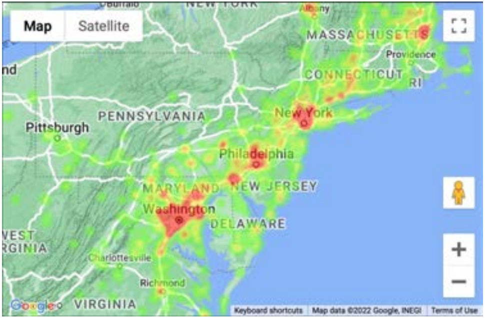


Heat map of the number of reported sightings

Image created from data taken from https://amsmeteors.org/2013/03/update-for-march-22-2013-northeast-fireball

Based on this, update your pdf for the meteor's direction, i.e. draw a new graph. Call the new pdf $f(\theta|x_1)$.
Estimate a 0.5-probability interval. How long is it?

Board question: Meteor! (c) Finer heat map.

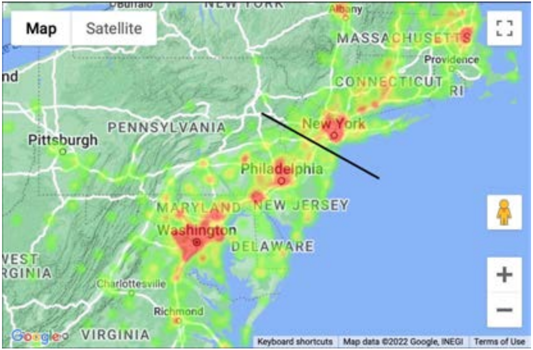


Heat map of the number of reported sightings

Draw a new pdf, $f(\theta|x_2)$, for the meteor's direction.

Draw a 0.5-probability interval. How long is it?

Discussion: Meteor! Actual direction.



(Approximate direction computed by amsmeteors.org referenced earlier.)

Discussion: how good is the data of the heat map for determining the direction of the meteor?

xkcd commentary

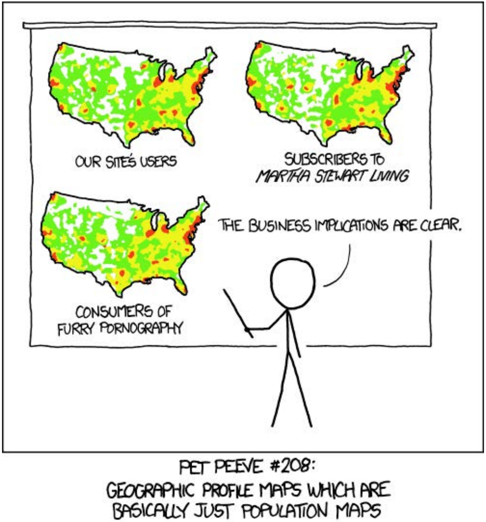


Image courtesy of [xkcd](https://xkcd.com). License: CC BY-NC.

`https://imgs.xkcd.com/comics/heatmap.png`