Tom Dienes
9.66 Pset 1
Prof. Tenenbaum

Problem 1

(i) (New Cover Story)
One day you find a bag of strange-looking coins lying on the sidewalk in front of a bank. **The bag contains exactly three coins.** You've never seen coins like these before and decide to examine them. They have recognizable "heads" and "non-heads" ("tails") sides; perhaps they're foreign coins? You decide to try flipping them and obtain the sequences below. Each sequence was generated by flipping a **random** coin from the bag. **After a coin is flipped, it is replaced in the bag to allow replacement.**

For each of the following sequences, please judge how likely you think the coin is to be a fair coin (tends to land heads half the time and tails half the time) or an unfair coin (tends to land on one side more often than the other). Use the 1-7 rating scale given below.

***No scale was attached to the pset, so I used a scale where 1 is certain unfair, 7 is certain fair, and 4 is unsure one way or the other

(Changes in bold)

I anticipate that the different cover story will give the prior assumption that there are exactly two unfair coins and one fair coin, since strings of all tails, all heads, and mixed exist.. I believe this will polarize the results, as there is less uncertainty

Old Cover	1	2	3	4	5	6	7	8	9
Participant	6	6	4	6	4	1	6	5	1
Me	6	6	5	5	5	2	6	5	1

New Cover	1	2	3	4	5	6	7	8	9
Participant	7	7	2	7	5	1	7	5	1
Me	7	7	3	7	6	1	6	6	1

ii) I do see systematic differences. Considering only the two from external participants, as I am biased, there are a few distinctions. Selena's standard deviation is 2.49, while Zoe's was just 1.94. This validates the assumption that the new cover story would lead to greater polarization.

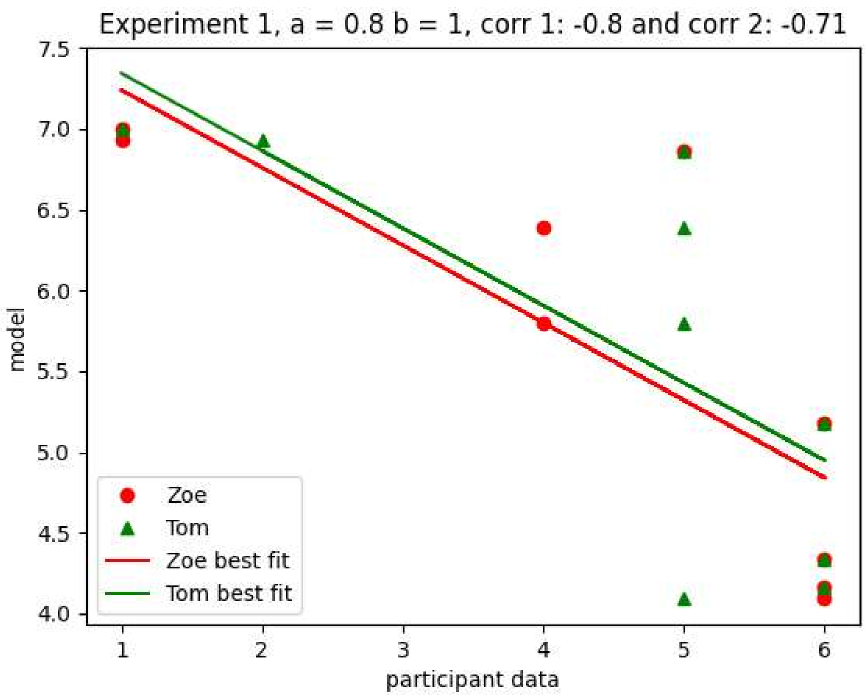
iii) For the most part, the changes were in line with my expectations. I expected that since there were three coins, people would be likely to assume there was one coin unfairly weighted towards tails, one coin unfairly weighted towards heads, and a fair coin. Under this belief, it seems more likely people would be confident enough to give a 1 or a 7 in more situations. This seems accurate, qualitatively. The old cover story produced 2 ones (confident unfair) and 0 sevens (confident fair). The new one produced 4 sevens and two ones.

Interestingly, the data didn't change the confidence much for trials 5, 8. This makes sense - while these were overwhelmingly heads, the presence of any tails shakes the previous idea that an entirely unfair coin was chosen, reducing our information back to the first cover story.

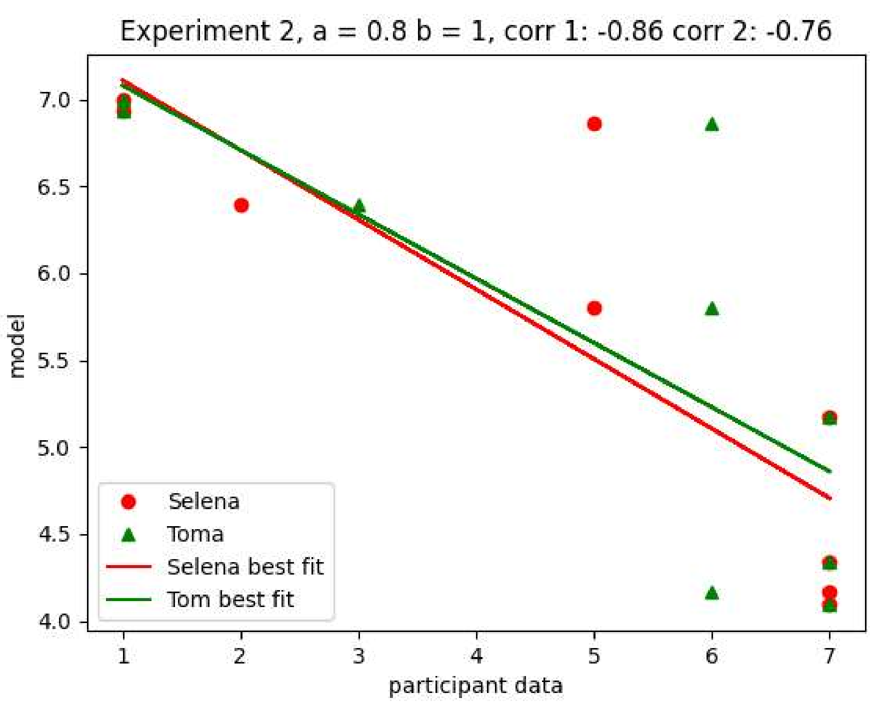
(b continued on next page)

(b)

(i)



Experiment 1, a = 0.8 b = 1, corr 1: -0.8 and corr 2: -0.71
model
participant data
Zoe
Tom
Zoe best fit
Tom best fit



Experiment 2, a = 0.8 b = 1, corr 1: -0.86 corr 2: -0.76
model
participant data
Selena
Toma
Selena best fit
Tom best fit

The average correlation of the first cover story is -0.76. The average correlation coefficient of the second experiment with the new cover story is -0.81.

(ii) I found the greatest success with (a=0.8, b = 1). In general, I noted that as I increased b towards 1 my correlation and thus goodness of fit increased.

(iii) Since a negative correlation value implies a stronger correlation (a stronger inverse), I calculated goodness of fit as the greatest R^2 (coeff of determination). This, similar to an absolute value, allows a more negative correlation to be seen as having more goodness of fit.

(iv) The model worked pretty well overall. The model assumes that randomness is largely determined by having an equal number of heads and tails. If we have all heads, as in sequence 7, it is much less likely to be perceived as fair. There weren't many systematic differences between the recorded data and the model. When using a=0.8, b =1, we see a correlation across both cover stories of ~-0.8, translating to an r^2 of ~0.64, which is pretty good. This suggests that the model accurately managed to portray the different data. In further support of the model, though the participants and I gave qualitatively different answers, after transforming the data through the model the best fit lines are parallel and nearly identical.

(c)

(i) The ratio $P(H_1)/P(H_2)$ is linearly dependent on $P(H_1)$, so decreasing $P(H_1)$ would cause the aforementioned ratio to decrease. This signifies that the participants would then have a higher likelihood of believing that the coins are unfair. This, in turn, would impact the model ratings over all sequences. If we were to increase $P(H_1)$, by the same token, the ratio would increase. This would then make the model think participants were likely to think the coins are fair, and make the model give each sequence a lower rating at all points.

(ii) Participants tended to be more likely to rate the coins as fair. The situations where someone was likely to rate a coin as unfair were generally like the final string of 20+ heads that is incredibly statistically unlikely. Because of this, I conclude that increasing $P(H_1)$ above 0.5 would better fit the participant's judgement.

(d)

(i) The hypothesis space only considers the comparison of the absolute quantity of heads to the absolute quantity of tails. This does not account for the ordering of the coins. For example, 'HHTHTTTHHT' and 'HHHHHTTTTT' have identical numbers of heads and tails, but the first sequence of numbers seems much more random. This makes sense only when an unfair coin has a certain weight to it skewing in one direction. If a coin could be unfair in the ordering like this, it would conflict with the hypothesis space.

(ii) As before, if we have a sequence with all heads (or tails) followed by all of the opposing side, the model would judge it as fair due to an equal number. A person would be less likely to do so. Additionally, with smaller numbers, a person would be likely to call "HH" fair, but the model would see only the differing numbers of heads and tails. Finally, if the sequence were alternating every other exactly "HTHTHTHT....", a person would (at sequence of a sufficient length) be likely to think unfair. The model, as before, would consider this fair.

(Problem 2 continued below)

Problem 2

(a)

2)

(a) Let H_1 be the hypothesis space of all mults of 10, and H_2 be all even nums

Then $P(y|h) = \frac{1}{|h|}$ iff $y \in h$

where y = elem, h = hypothesis

$P(H_1) = P(H_2) = \frac{1}{2}$ where data $= \{10, 70, 30\}$

$$p(h|\text{data}) = \frac{P(\text{data}|h) \cap P(h)}{\sum_{h' \in H} P(\text{data}|h') \cap P(h')}$$

$$P(H_1|\text{data}) = \frac{P(\text{data}|H_1)\, P(H_1)}{P(H_1)\, P(\text{data}|H_1) + P(H_2)\, P(\text{data}|H_2)}$$

$$= \frac{\left(\frac{1}{10}\right)^3 \times \frac{1}{2}}{\left(\frac{1}{10}\right)^3 \times \frac{1}{2} + \left(\frac{1}{50}\right)^3 \frac{1}{2}}$$

$$= \boxed{\frac{125}{126}}$$

$$P(H_2|\text{data}) = 1 - P(H_1|\text{data})$$

$$= 1 - \frac{125}{126} = \boxed{\frac{1}{126}}$$

(b)

(b) $P(y \in c \mid D) = \sum_{h \in H} P(y \in c \mid h)\, P(h \mid D)$

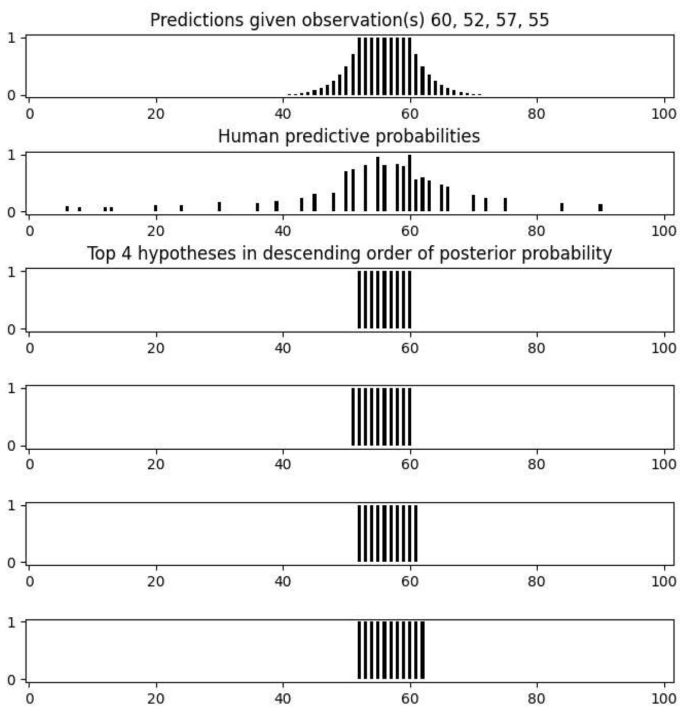
$= \cancel{P}\ 1 \times \frac{125}{126} + 1 \cdot \frac{1}{126} \boxed{= 1}$

Qualitatively, 40 is both even and a multiple of 10.

(c)

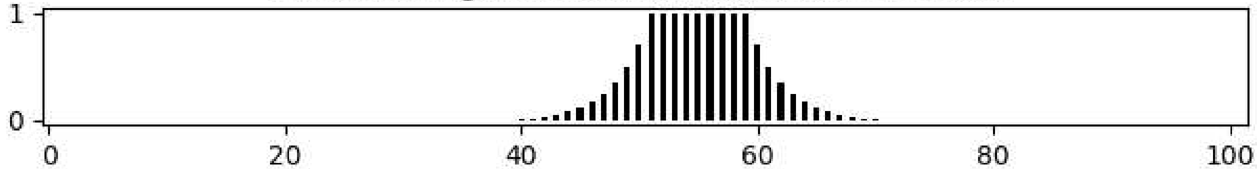
Code attached

(d)

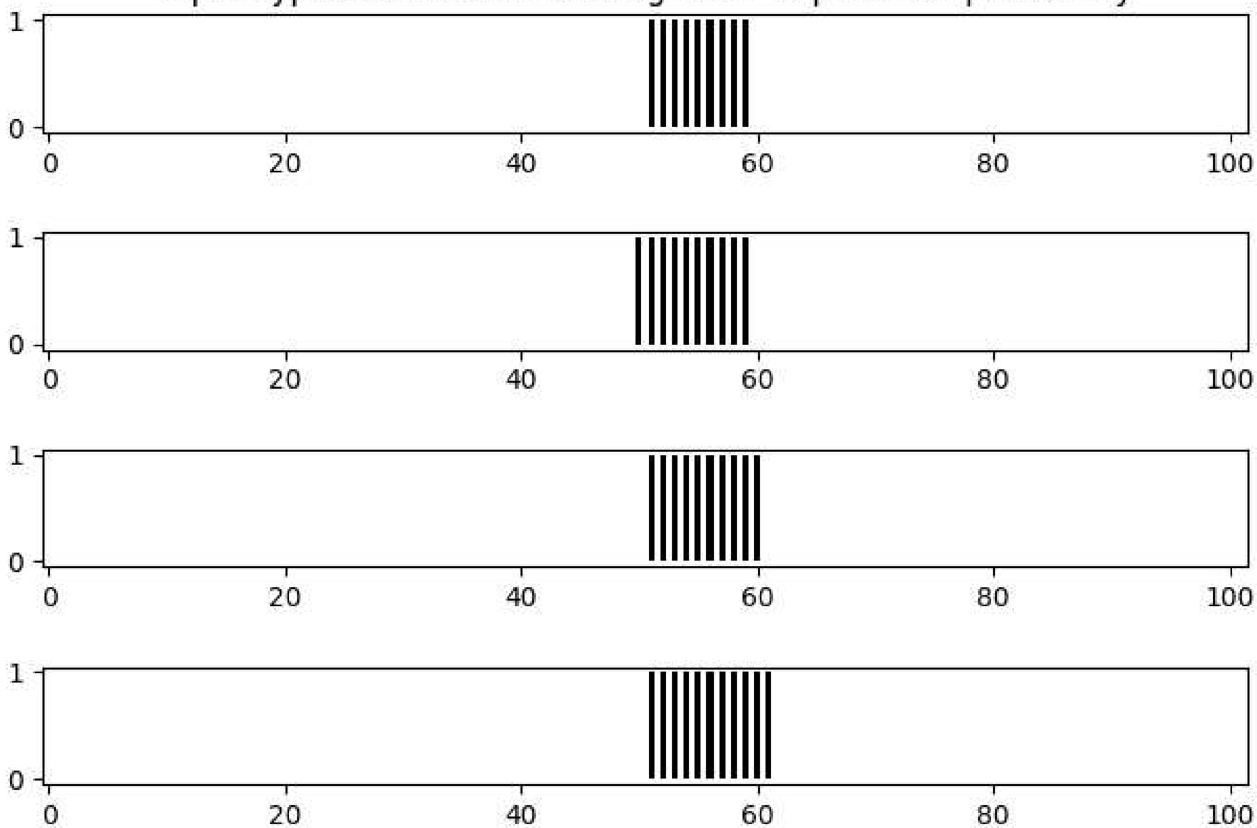


(i)

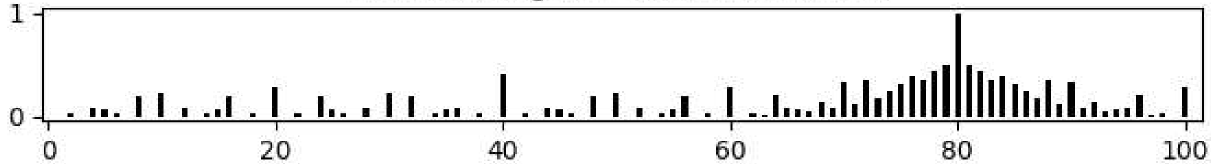
Predictions given observation(s) 59, 51, 56, 54



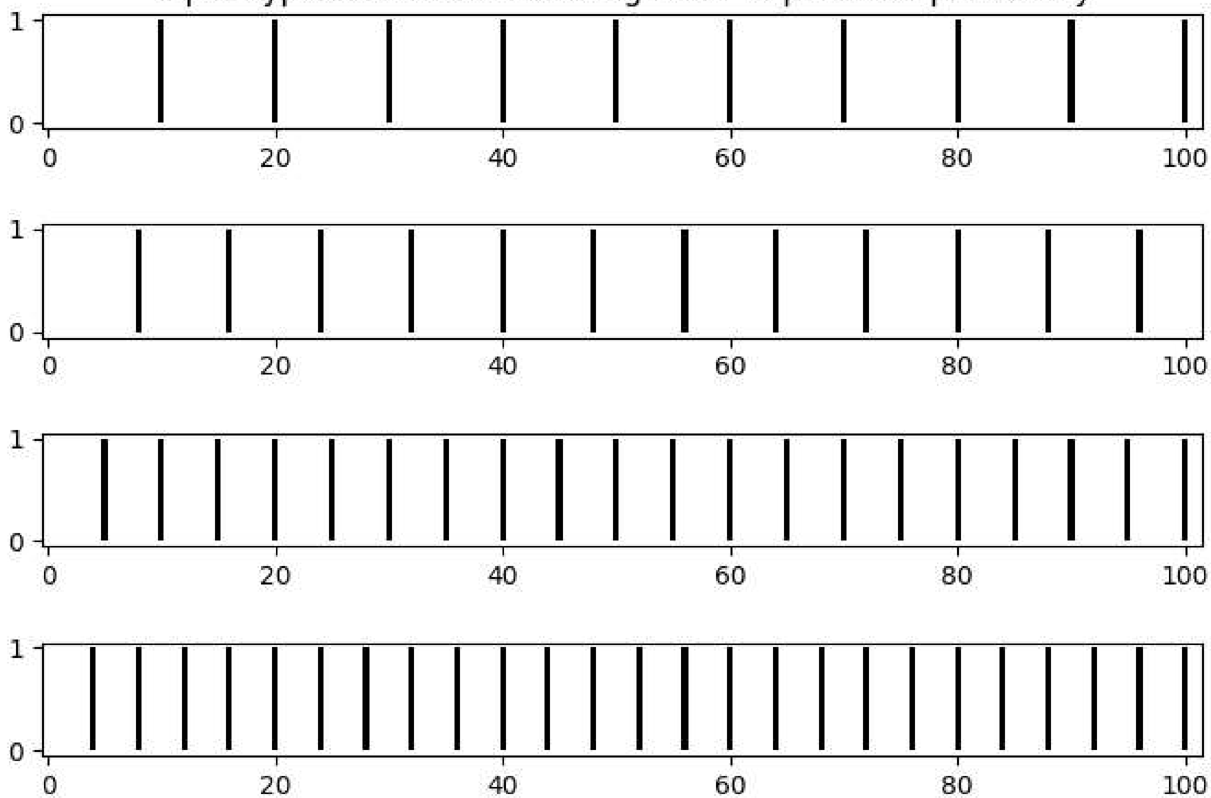
Top 4 hypotheses in descending order of posterior probability

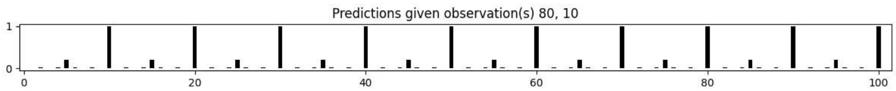


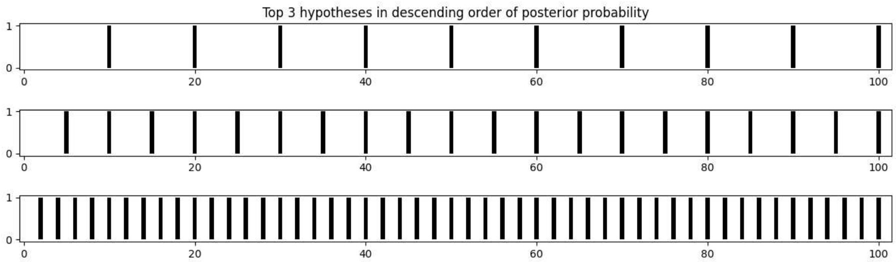
Predictions given observation(s) 80

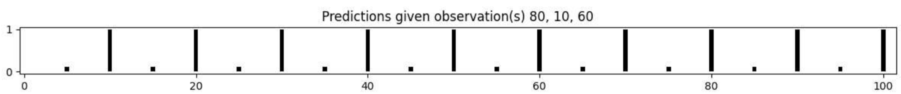


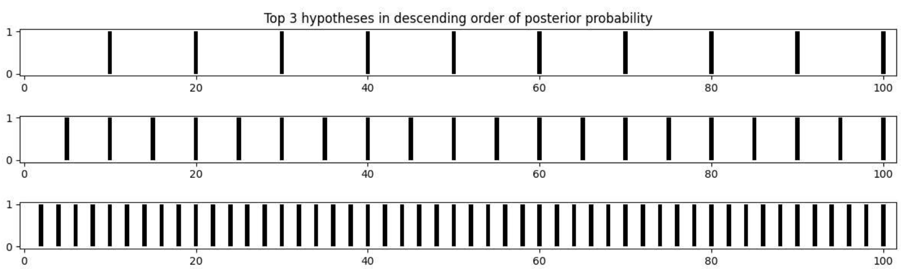
Top 4 hypotheses in descending order of posterior probability

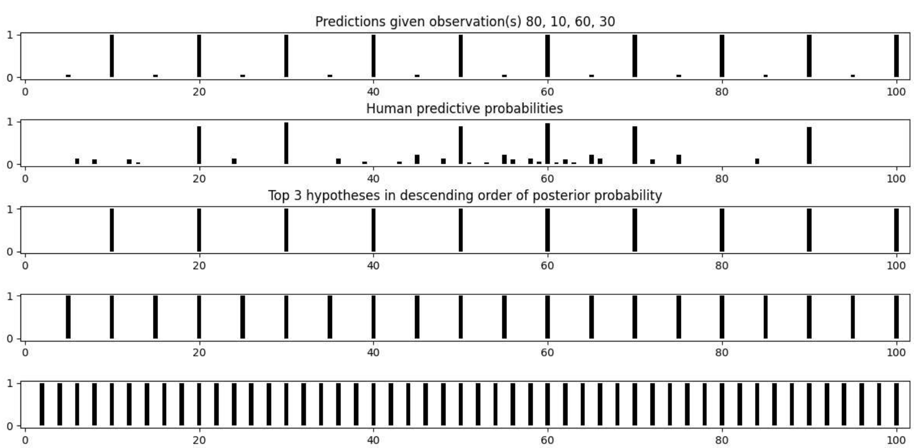






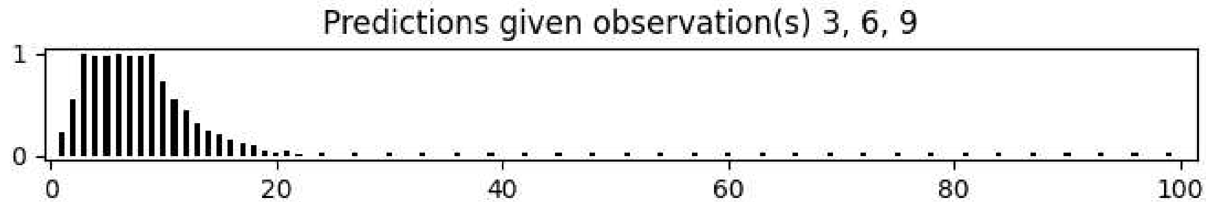


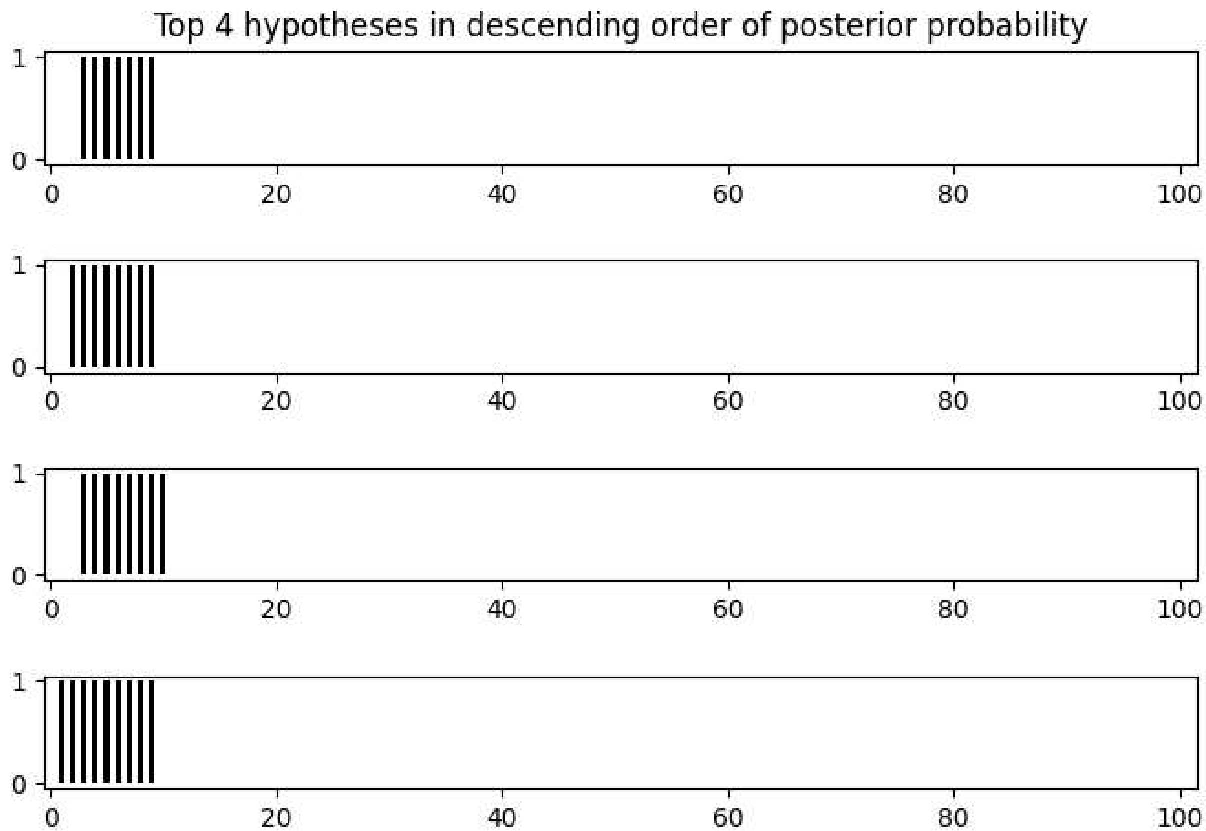




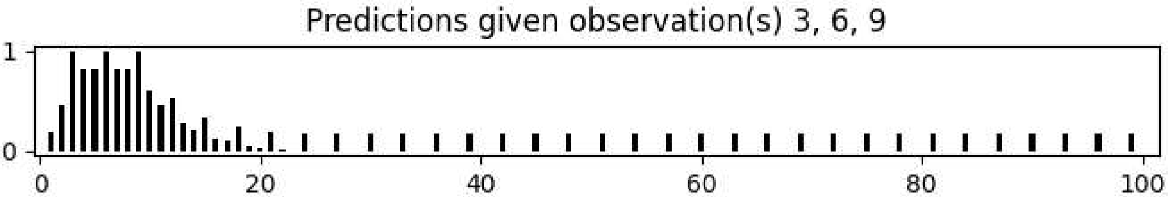
In the first dataset of just [80], we only have one multiple of 10. By the fourth, there are four ([80,10,60,30)]. As more multiples of 10 are added to the dataset, new data changes the predictive distribution to skew it towards the multiple of 10 hypothesis. This makes a lot of intuitive sense - as there are more consecutive multiples of 10 in our dataset, it becomes increasingly likely that n*10 is our hypothesis space.

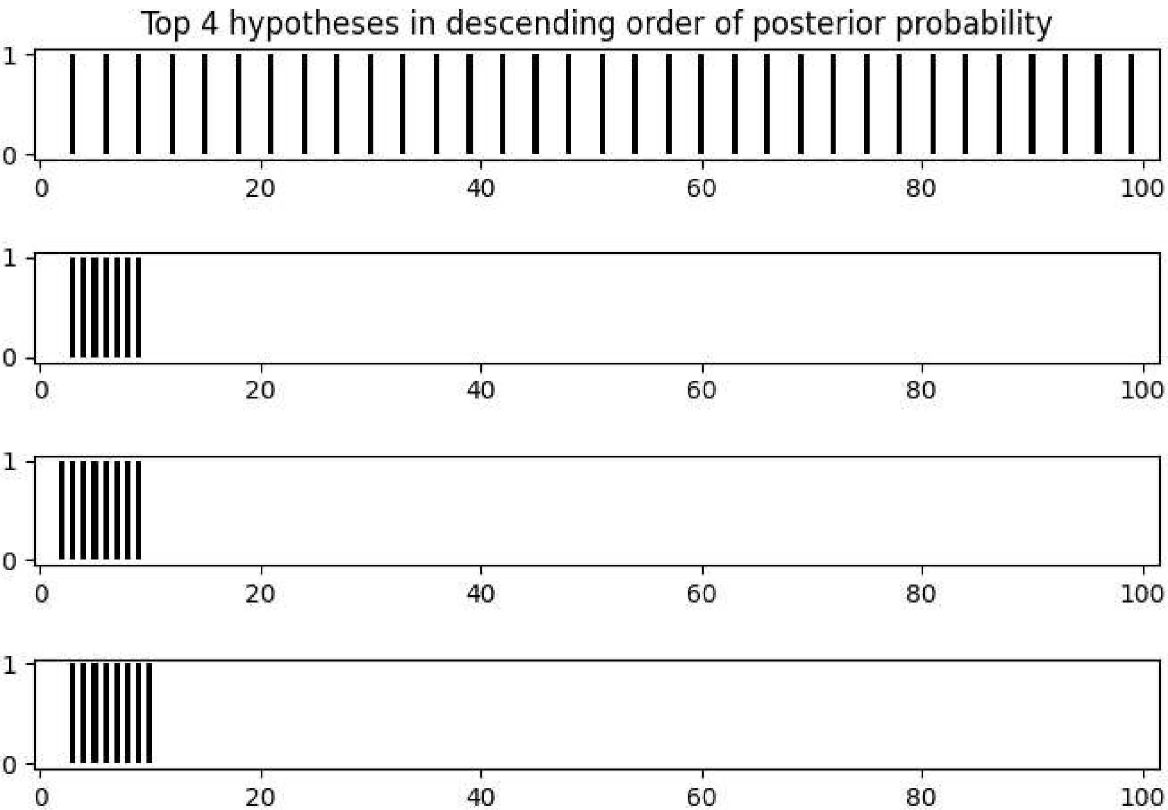
(iii)



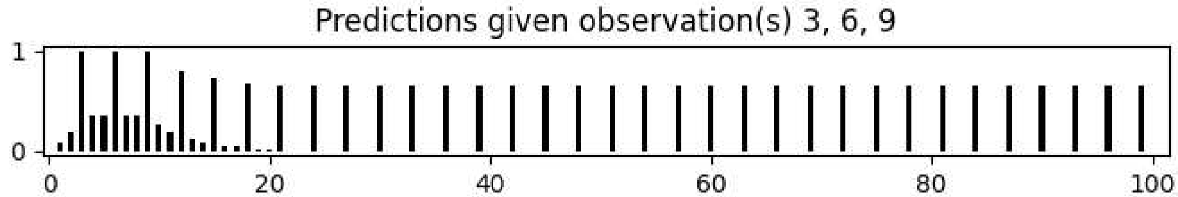


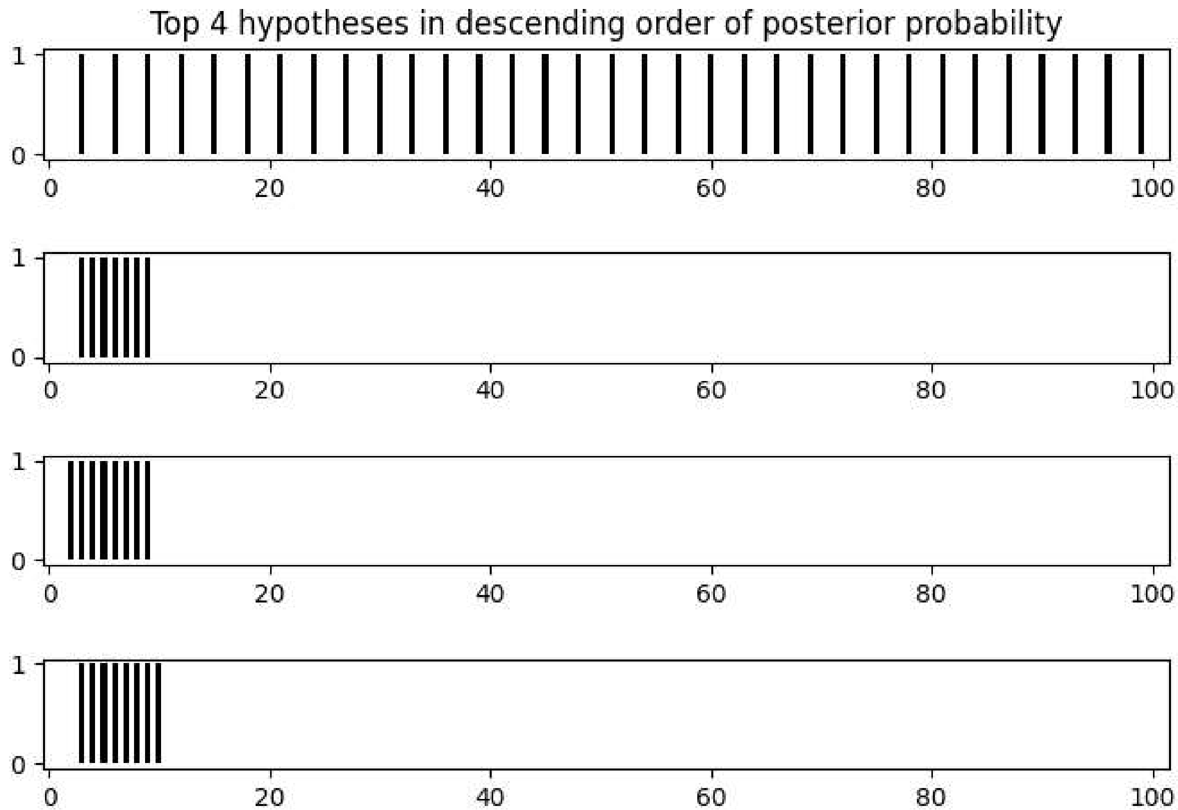
(priors = (0.9, 0.1))


Predictions given observation(s) 3, 6, 9
1
0
0
20
40
60
80
100


Top 4 hypotheses in descending order of posterior probability
1
0
0
20
40
60
80
100

(priors = (0.5, 0.5))





(priors = (0.1, 0.9))
It seems as if a smaller interval prior indicates that some pattern has an increased likelihood of being included in the concept, and that all numbers outside of this pattern are not in this concept. This can be observed by the spike in high predictive values at some numbers while other predictive values are much lower. When the interval prior is larger, all numbers are evened out such that numbers within the range of the interval itself have a significantly higher predictive value. Qualitatively, this is due to a bias towards the data from the interval itself. Overall, a smaller interval prior makes prediction depend more on an external pattern believed to exist, and a larger interval prior makes the prediction depend more on the numbers themselves and the range in which they exist.

(iv) When creating these numbers, I definitely had the pattern in mind of multiples of three. Based off of that, the prior of (0.9, 0.1) seems to be the worst, as it is the only one that produces exclusively intervals (from 0-9 or 3-9) as the most likely hypotheses. When compared to the data with human responses, it can really depend. It doesn't seem like there is one prior setting that works the best for every possible dataset. For these, the given values of (0.8, 0.2), or even (0.7, 0.3) seems to work pretty well. However, people are pretty able to recognize if a sequence is a pattern or interval across a wide range of priors. That said, people will start to fail relatively quickly if the dataset has a very low number of datapoints. With just one number, a person would be biased towards thinking [17] is a member of a range and not a multiple, although there is little info backing that up.

(e)

(i) As far as Marr's levels are concerned, I believe that the number game is largely on the computational side. Really, this game takes in datapoints and then tries to explain how humans will analyze that data and then predict the future sequence. The number game focuses on this sequence analysis and then captures information based on whether or not those sequences are more likely to belong to an underlying mathematical pattern or if they are randomly chosen from an interval range. The number game does also have some algorithmic aspects to it, as it does in many ways explain how to solve identified computational problems. The number game does not get specific enough to be implementational. The game captures information about how people analyze patterns and also how people change what hypothesis they believe a dataset belongs to as more information is added.

(ii) I think that the number game is slightly ecologically relevant. The game allows us to observe how humans analyze sequences of numbers to try to determine what the priors are, as well as how humans are able to update their initial conceptions for those sequences of data when more information is added. This feels like a simplified version of things we experience as humans every day, where we are forced to come up with preconceived notions for something and then adjust our opinion as more data is added. The number game, clearly, is rudimentary and only deals with highly specific formats of data restricted to a pattern or a interval. In real life, outside the lab, we will never run into this exact situation. I claim it is still ecologically relevant. For example, the number game could be used to observe whether length of time a prior is held impacts the ease with which a participant changes the hypothesis. This is directly applicable to reality. To be more ecologically relevant, experiments based off the number game would need to try to incorporate a more complex sequence of data to better represent the massively complex web of data that is considered with real-life human cognition.

(iii) I think choosing between a pattern of numbers or some range is pretty intuitive, and what most people are going to naturally choose. So, I think people playing the number game are likely to pick a similar one to the hypothesis space discussed above. There are certainly other options - for example, if the numbers are restricted to 0-26, instead of being intervals, there could be some alphanumeric pattern when translated to the index in the alphabet. Additionally, geometric or abstract patterns are also conceivable. Finally, it is possible to construct a hypothesis space from real life events. (eg. number of a year in which a democrat was elected). That said, I think that these are possibilities and not the most likely idea to be conceived, and so choosing between a pattern and interval is the natural, intuitive hypothesis space.

Problem 3 continued below

Problem 3
(a)

$$P(t_{total}) = t_{total}^{-\gamma}$$

$$P\left(\frac{t_{total}}{t_{obs}}\right) = \frac{P(t_{total})P(t_{obs}|t_{total})}{P(t_{obs})}$$

$$P(t_{obs}) = \int_0^\infty \left(P(t_{obs}|t_{total})\, p(t_{total})\right) dt_{total}$$

$$P(t_{obs}|t_{total}) = \cancel{\int_0^\infty P(obs|total)}$$

$$= \frac{1}{t_{total}} \text{ when } t_{obs} < t_{total}$$

we assume tobs is chosen randomly from within the interval P(total)

we calculate the posterior median as x where $P(\frac{t_{total}}{t_{obs}})$ has $\frac{1}{2}$ vals above and below

Combining everything above

$$P(t_{total}|t_{obs}) = \frac{(t_{total}^{-\gamma})(P(t_{obs}|t_{total}))}{\sum (t_{total}^{-\gamma}\, dt_{total}) P(t_{obs}|t_{total})}$$

$$= \frac{t_{total}^{-\gamma} \frac{1}{t_{total}}}{\sum_{t_{obs}} \frac{1}{t_{total}} t_{total}^{-\gamma}\, dt_{total}}$$

$$= \frac{t_{total}^{-(\gamma-1)}}{\sum t^{-(\gamma-1)}\, dt_{total}}$$

$$= \frac{t_{total}^{-(\gamma-1)}}{\lim_{x\to\infty} -\frac{1}{\gamma} t_{total}^{-\gamma} \Big|_{t_{obs}}^{x}}$$

$$\boxed{= \gamma \frac{t_{obs}^{\gamma}}{t_{total}^{\gamma+1}}}$$

(b)

(b) Find t^* s.t. $P_r[t^* > t_{total} | t_{obs}] = 0.5$

Suppose: $P_r[t^* > t_{total} | t_{obs}]$

$$= P_r(t^* > t_{total}) \cdot P(t_{total} | t_{obs})$$

$$= \int_{t_{obs}}^{t^*} P(t_{total} | t_{obs}) \, dt_{total}$$

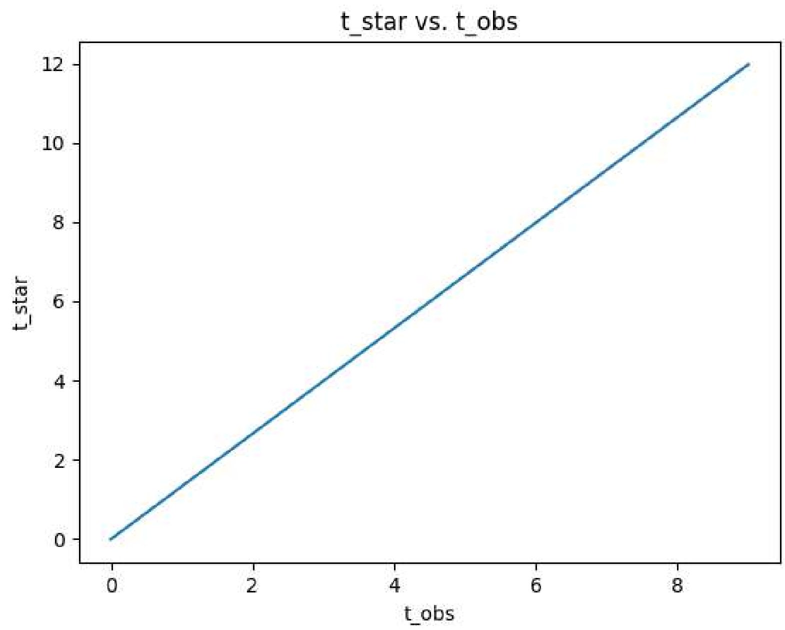
work in (a) $$= \int_{t_{obs}}^{t^*} \gamma \frac{t_{obs}^{\gamma}}{t_{total}^{\gamma+1}} \, dt_{total}$$

$$= 1 - \frac{t_{obs}^{\gamma}}{t_*^{\gamma}}$$

and this must $= 0.5$

so $$0.5 = 1 - \frac{t_{obs}^{\gamma}}{t_*^{\gamma}}$$

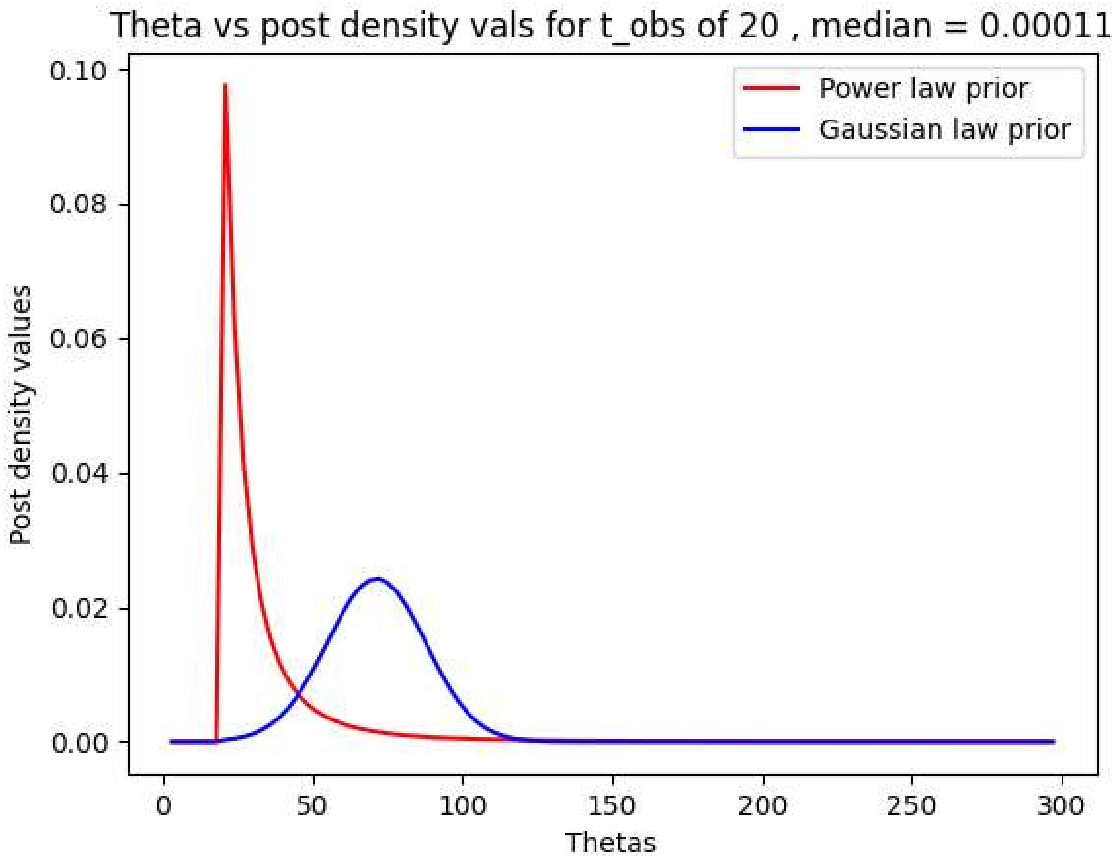
$$\Rightarrow t^* = \frac{t_{obs}}{\left(\frac{1}{2}\right)^{\frac{1}{\gamma}}} = \boxed{2^{\left(\frac{1}{\gamma}\right)} t_{obs}}$$

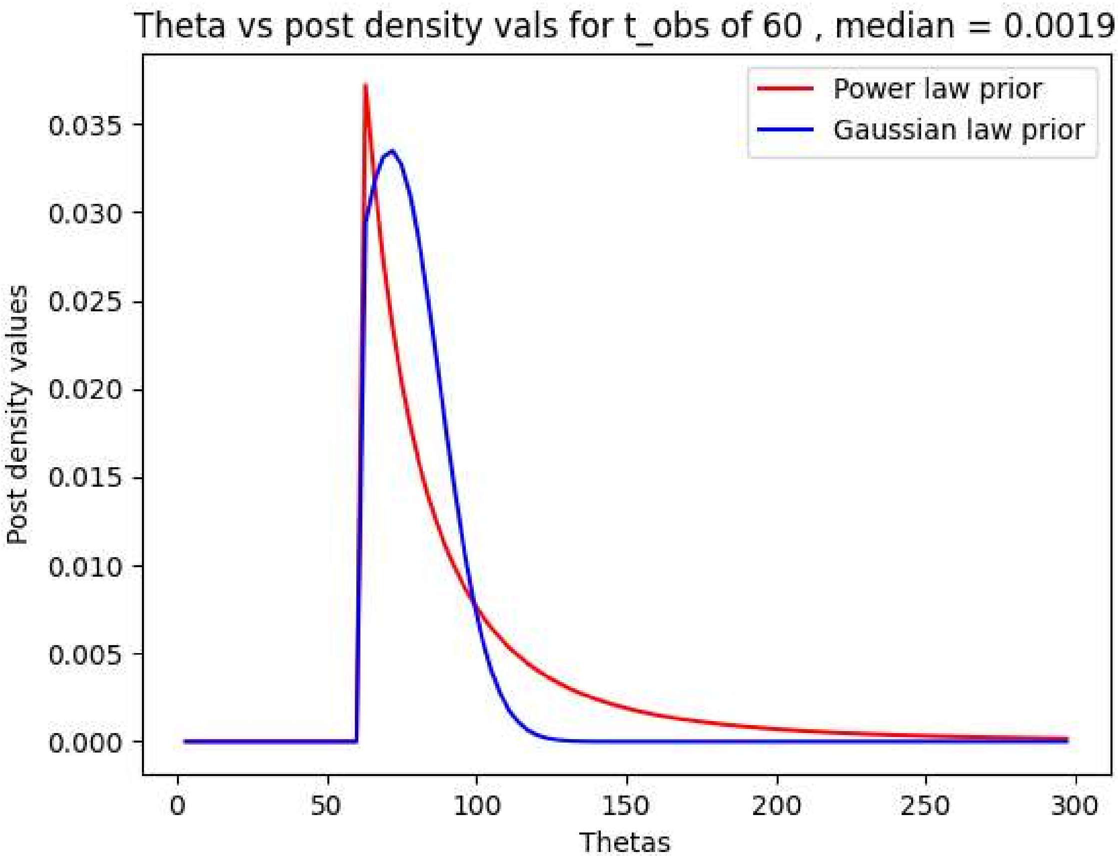


(c) (d) (e) Code attached

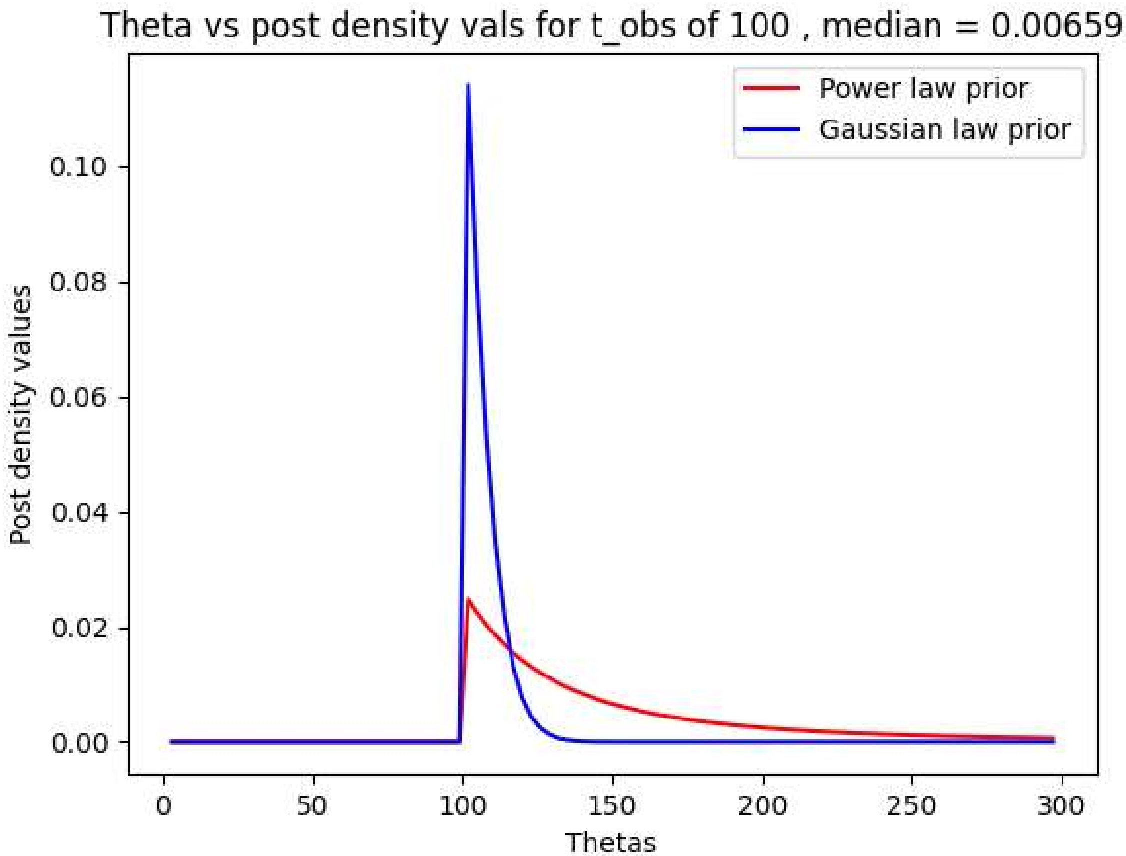
(f)

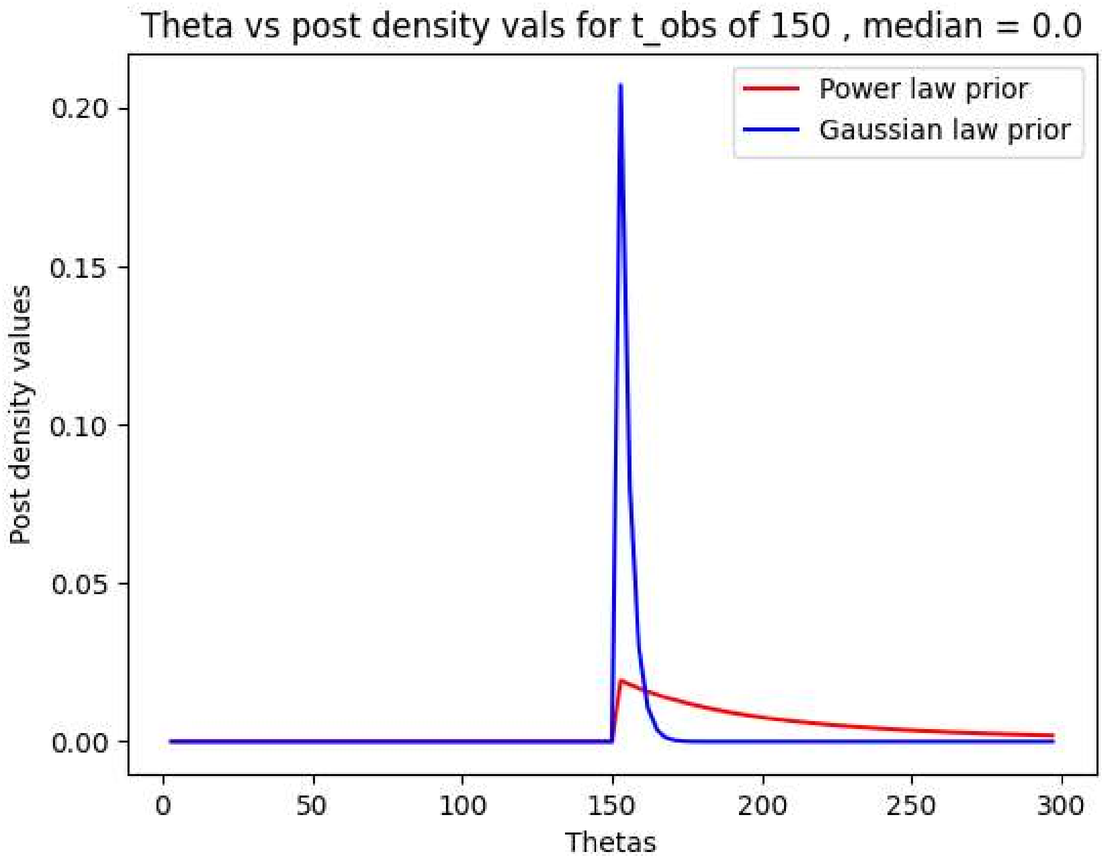
Below are attached the resulting graphs of posterior density values vs thetas with Power Law and Gaussian priors for the different range of observed values [20, 60, 100, 150]. As you can see, as t_{obs} increases, the median predictive value increases across both the Power Law and the Gaussian priors. Intuitively, this makes sense. As we observe higher values, we will start to predict larger values for the total. However, the Gaussian prior is much more loosely banded for larger theta values than for the Power Law. The Power Law, instead, prefers to predict tightly around a small theta range. The Gaussian prior appears to increase much slower as t_obs grows greater than 100. Intuitively, this is verified as numbers greater than 100 are outside of the range of what we could consider believable. Interestingly, from the final graph, we observe that the Power Law median grows linearly with t_obs, while the Gaussian prior drops off as t_obs increases. This is consistent with the slow growth we saw over 100.

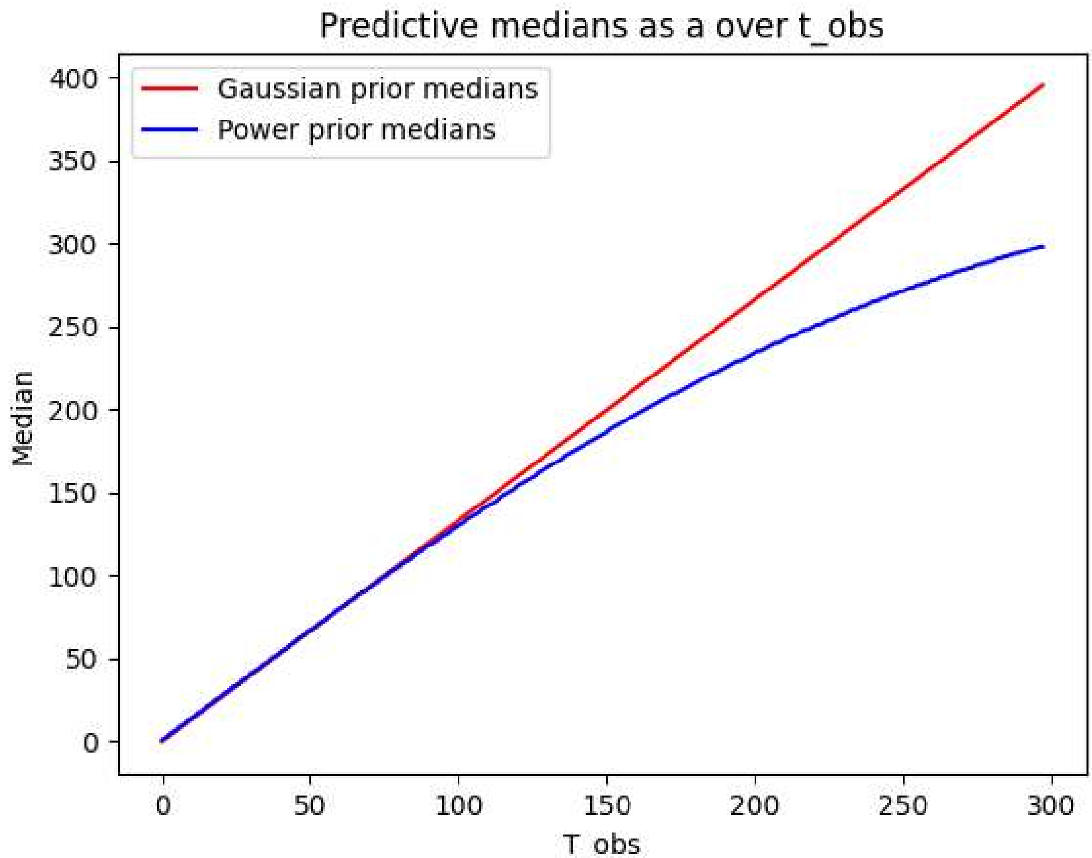

Theta vs post density vals for t_obs of 20 , median = 0.00011
Power law prior
Gaussian law prior
Post density values
0.10
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
250
300
Thetas



Theta vs post density vals for t_obs of 60 , median = 0.0019
Power law prior
Gaussian law prior
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Post density values
0
50
100
150
200
250
300
Thetas


Theta vs post density vals for t_obs of 100 , median = 0.00659
Power law prior
Gaussian law prior
Post density values
0.00
0.02
0.04
0.06
0.08
0.10
0
50
100
150
200
250
300
Thetas


Theta vs post density vals for t_obs of 150 , median = 0.0
Power law prior
Gaussian law prior
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
250
300
Thetas
Post density values



Predictive medians as a over t_obs
Gaussian prior medians
Power prior medians
Median
T_obs
0
50
100
150
200
250
300
350
400