AGH

Analysis of fatal accidents on Mount Everest 1990 – 2019

Data Analytics Project

Authors: Jakub Majcher, Rafał Skrzypek

Cracow, 2023

1. Problem formulation

Statistics from Himalayan expeditions are available for public review. The following data set has information on particular expeditions, members, deaths for each peak in Himalayan Mountain Chain. Those data have been collected since 1920. Every expedition contains information about status, mountaineers, heights, oxygen use or cause of death.

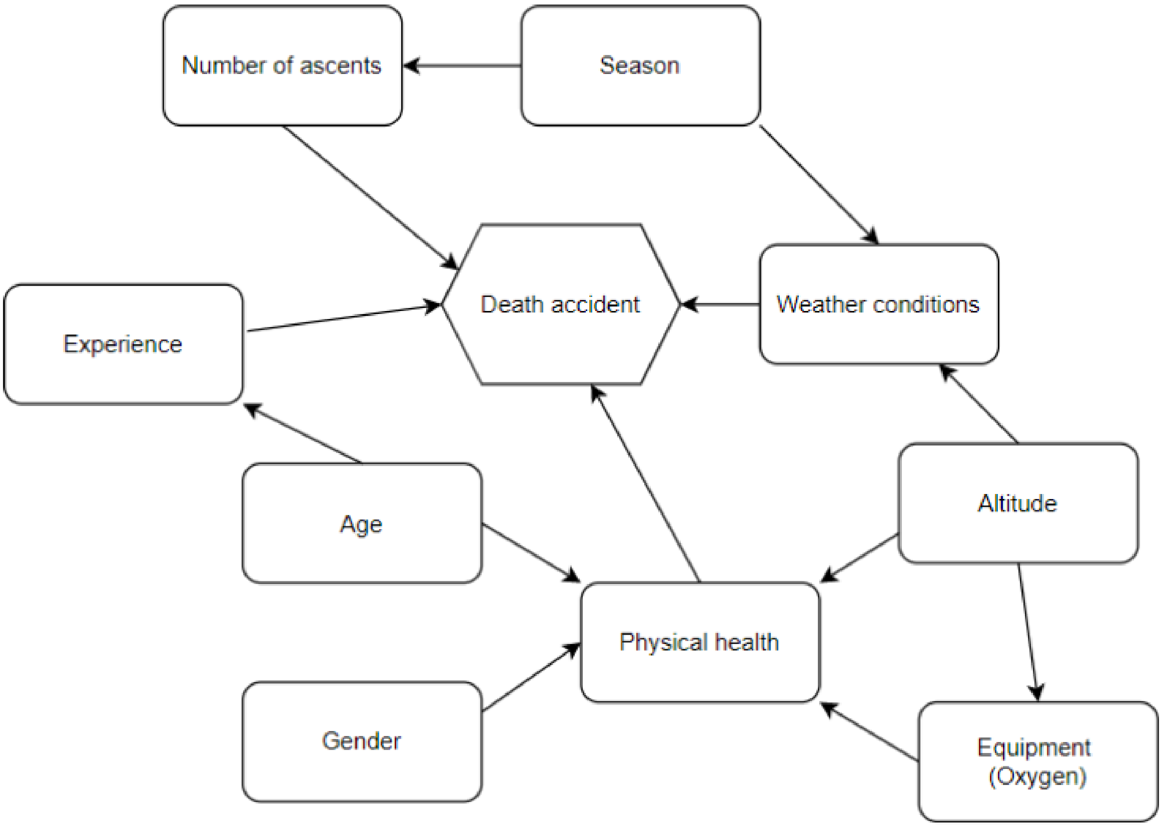
The main goal of this project is to describe the death rate and investigate the correlation between fatal accident and number of members taking part in expeditions.

The created models may help to forecast possible accidents in the future and help to prevent them. Additionally, this model can be used to improve safety during expeditions by contribution to educating an raising awareness about conducting safe expeditions in Himalayas.

Database is available on https://www.himalayandatabase.com/index.html

DAG

Directed acyclic graph created based on preprocessed dataset.



Possible confoundings

Forks

- Physical health and weather conditions have common cause of altitude,
- Physical health and Experience have common cause of age,
- Number of ascents and weather conditions have common cause of season.

Colliders

- Physical health is influenced by gender, age, equipment, altitude,
- Weather conditions are influenced by altitude ad season,
- Death accident is influenced by experience, number of ascents, weather conditions and physical health.

Pipes

- Gender is transmitted through physical health to fatal accident

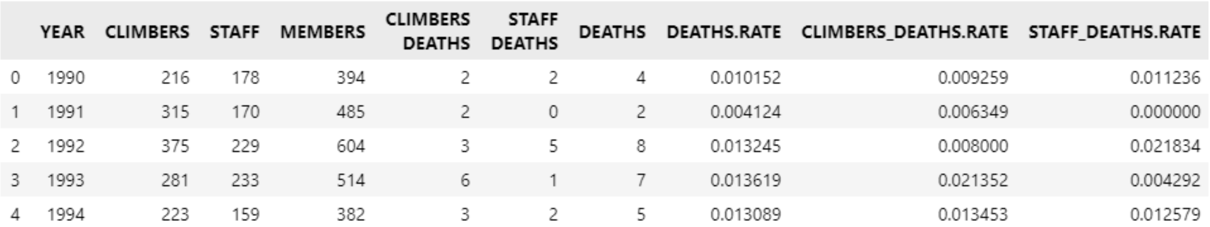
2. Data preprocessing

The Himalayan database contains a lot of information about expeditions, climbers, staff members, summits, weather conditions, oxygen use, cause of deaths in Himalayan Mountain Chain. To prepare data for modelling only useful data were chosen. We decided to analyze fatal accidents on Mount Everest between years 1990-2019 because during there was a lot of ascents and information. We rejected year 2020 and next because of the pandemic. During this period, it was hard to arrange expeditions.

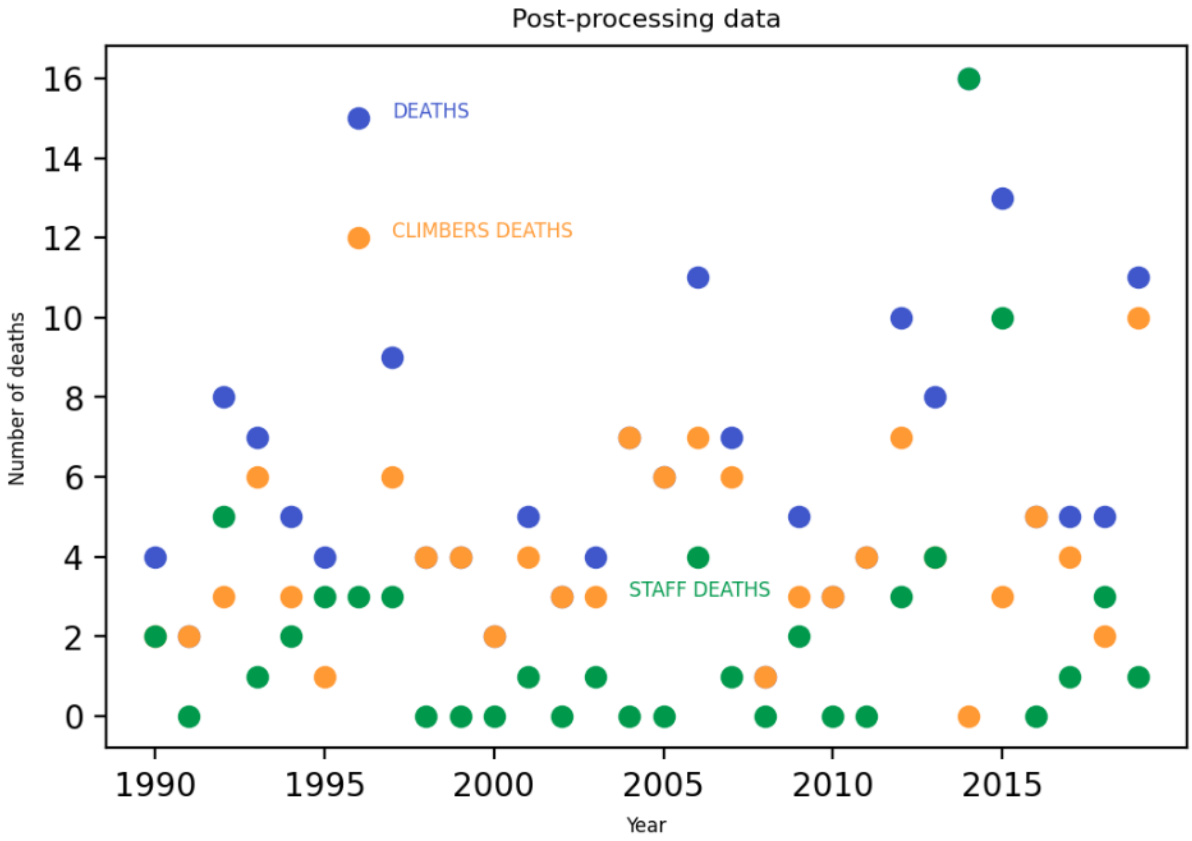
From the database we have chosen useful information and prepared EVEREST.xlsx file containing:

- Year
- Number of climbers
- Number of staff members
- Death cases of climbers
- Death cases of staff members

And prepared data frame visible below:

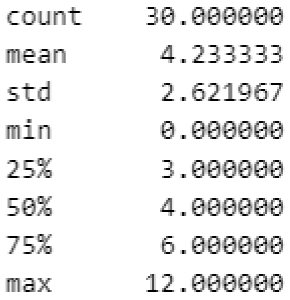


	YEAR	CLIMBERS	STAFF	MEMBERS	CLIMBERS DEATHS	STAFF DEATHS	DEATHS	DEATHS.RATE	CLIMBERS_DEATHS.RATE	STAFF_DEATHS.RATE
0	1990	216	178	394	2	2	4	0.010152	0.009259	0.011236
1	1991	315	170	485	2	0	2	0.004124	0.006349	0.000000
2	1992	375	229	604	3	5	8	0.013245	0.008000	0.021834
3	1993	281	233	514	6	1	7	0.013619	0.021352	0.004292
4	1994	223	159	382	3	2	5	0.013089	0.013453	0.012579



Each year differs in terms of fatal accidents. In some years, only climbers were affected by accidents, while in others, only staff members were involved. The main cause of such incidents is the changing weather conditions on Mount Everest, which lead to blizzards and avalanches. In our project we decided to analyze only climbers deaths.

Climbers deaths described:



```
count    30.000000
mean      4.233333
std       2.621967
min       0.000000
25%       3.000000
50%       4.000000
75%       6.000000
max      12.000000
```

3. Models

Two Different Models Specified:

- The Poisson Model: The Poisson model assumes that the number of deaths follows a Poisson distribution, which is a discrete probability distribution used for events that occur at a constant average rate over a fixed interval of time or space.
- Modified Poisson Model: The modified Poisson model incorporates an additional parameter that is dependent on the number of climbers each year. This parameter allows for the possibility that the average rate of deaths may vary depending on the number of climbers.

Difference Between Two Models Explained:

- The Poisson Model: In the basic Poisson model, the parameter represents the average rate of deaths over a specific interval. It assumes that the rate remains constant regardless of any external factors such as the number of climbers.
- Modified Poisson Model: The modified Poisson model includes an additional parameter that captures the potential relationship between the number of climbers and the average rate of deaths. This acknowledges the possibility that the risk of death may increase or decrease with the number of climbers present on Mount Everest each year.

Justification for the Difference in Models:

- The inclusion of the additional parameter in the modified Poisson model is justified because it accounts for a potential relationship between the number of climbers and the risk of death. It allows for a more nuanced analysis by considering how the average rate of deaths may vary based on the number of climbers present.

Sufficient Description of the Models:

- The Poisson Model:

$$y_i \sim Poisson(\lambda)$$

λ is the parameter representing the average of deaths.

- Modified Poisson Model:

$$y_i \sim Poisson(\theta n)$$

λ can be decomposed to intensity θ and exposure n which is the number of climbers who went on expeditions in a given year

4. Prior first model

The mean of fatal accidents of climbers per year is approximately 4, but there are years when this number is much higher (e.g., 12 in 1996). Therefore, we assume that the probability of having 4 times the number of accidents observed in 1996 is 1%.

To find lambda, which is the average in the Poisson distribution, lambda is the mean value of events, and the square root of lambda is the standard deviation. So, to find lambda for which approximately 99% of the probability density of the distribution will fall within the range up to 48, we need to solve the equation:

$$\lambda + 3 * \sqrt{\lambda} \approx 48$$

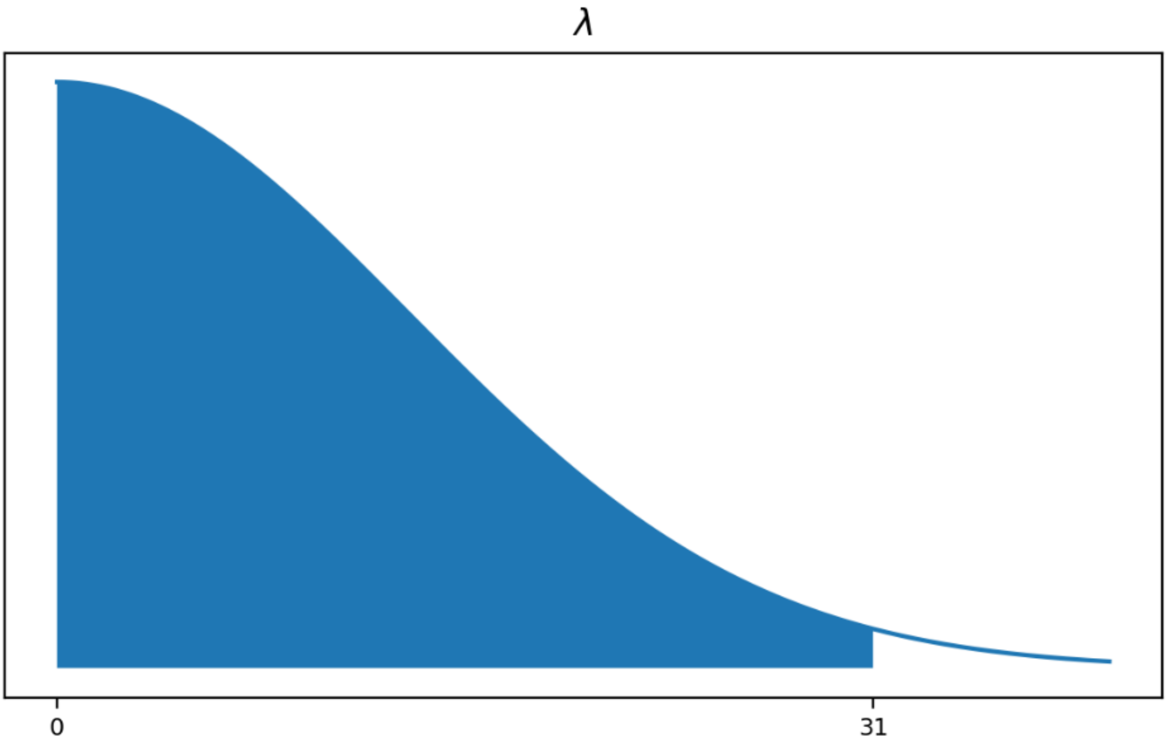
Solution for λ > 0:

$$\sqrt{\lambda} \approx 5.589$$

$$\lambda \approx 31$$

We know that for a normal distribution with a mean of 31 and a standard deviation of the square root of 31, there is a 1 percent chance that the number of deaths in a given year will exceed 48. The next step is to model the prior distribution for lambda. To do this, we will use a half-normal distribution for which 99 percent of the density will be less than the previously calculated threshold value of lambda, which is 31.

The standard deviation for such a distribution using the Newton's method is 13.325.

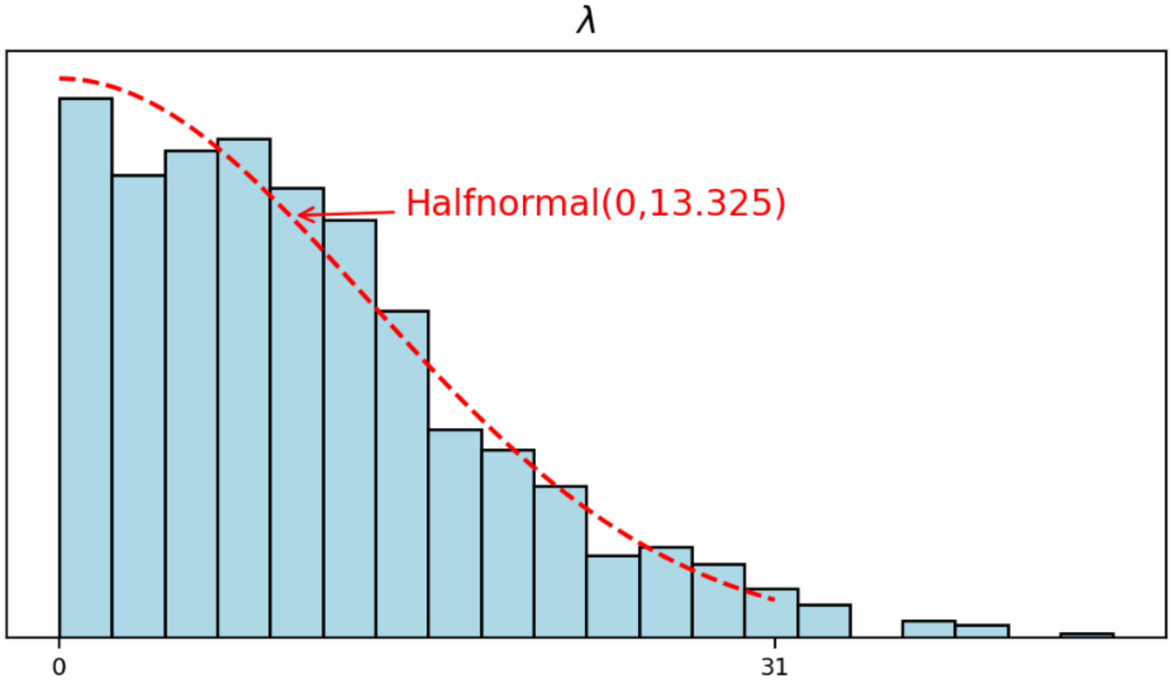


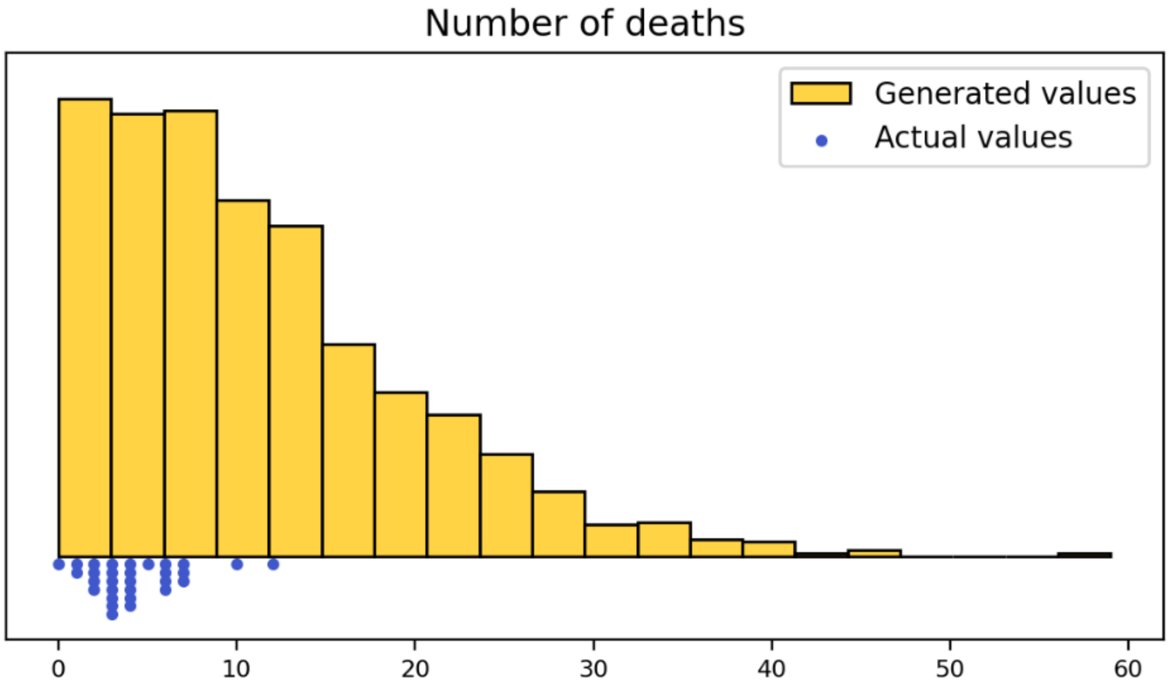
Prior predictive checks

In this model Poisson distribution is used to generate samples based only on prior. We are using absolute values of normal distribution with mean value = 0 and calculated before standard deviation = 13.325. This allows to avoid sampling errors.

```
data {
    int M; //Number of years analyzed
}
generated quantities {
   real lambda =fabs(normal_rng(0,13.325));
```

```
    int y_sim[M];
    for (k in 1:M){
     y_sim[k] = poisson_rng(lambda);
    }
}
```





Generated λ parameter is consistent with halfnormal prior.

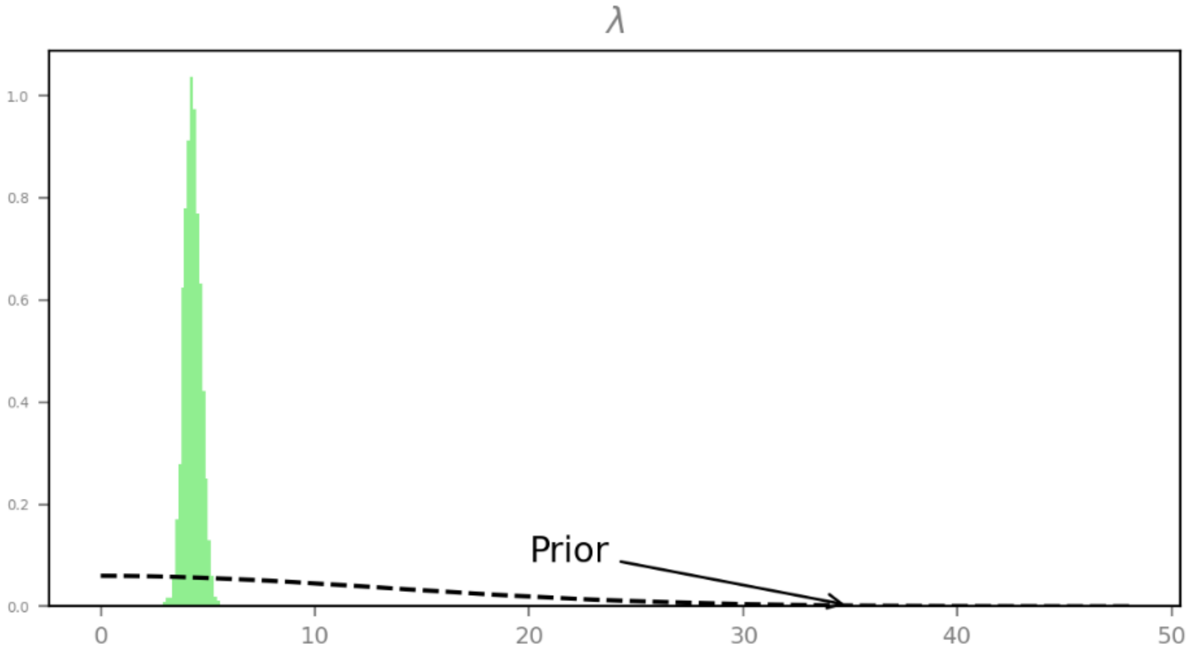
Generated number of deaths is located in the most probable area so prior is consistent with it. The disadvantage of the prior might be the fact that the highest probability of fatal accident expects more samples with low values, but there are only 1 case where was only 1 death within a year and 2 cases where there 2 and 3 deaths. But probability of higher number of fatal accidents decreases, which is correct. Tail is filled up to 48 in accordance with the assumption.

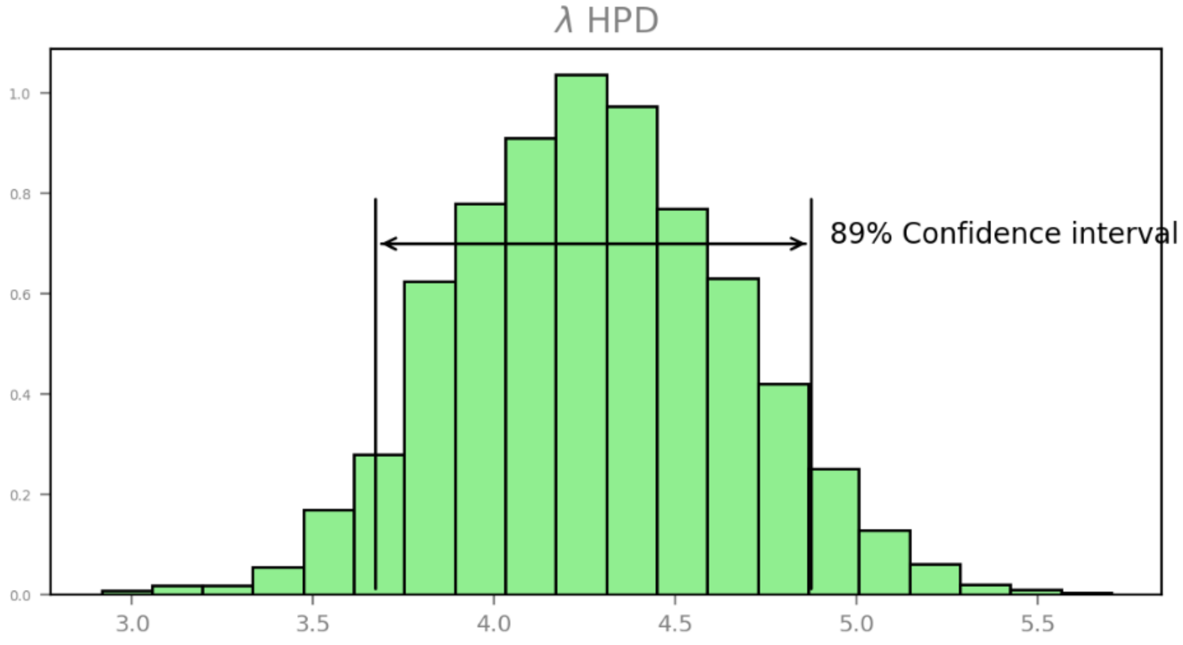
5. Posterior first model

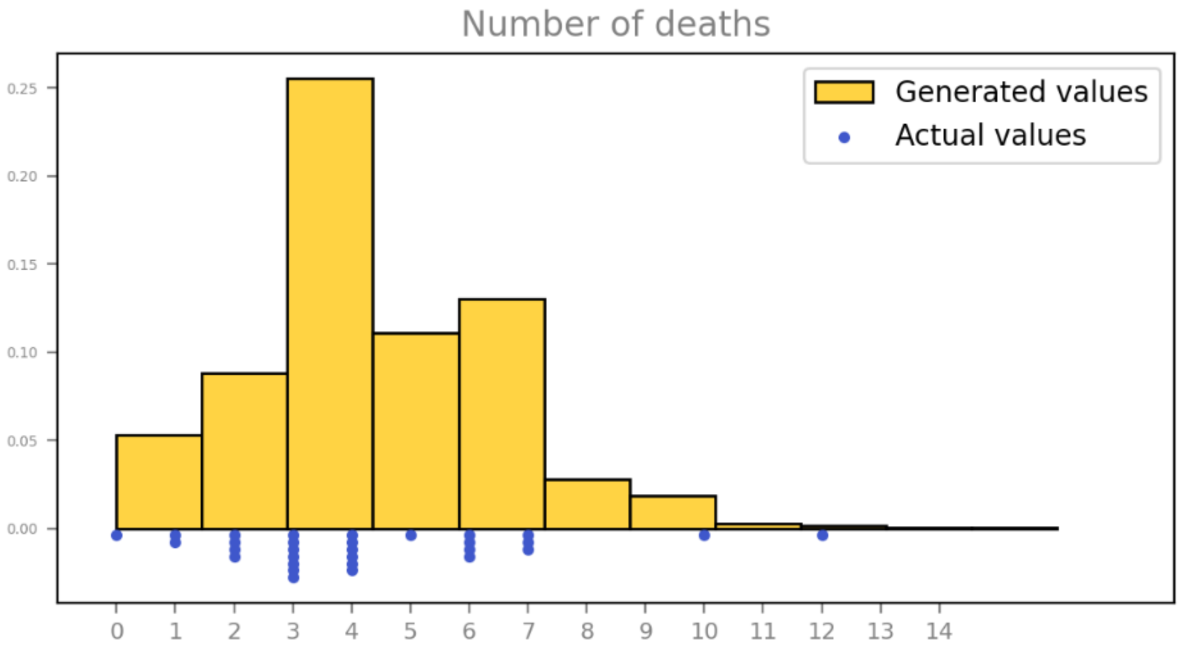
To infer the value of lambda the model containing:

- data block with actual number of fatal accidents within a year,
- parameter block with the value of λ with enforced constraints to get halfnormal distribution of λ in the model
- model block with likelihood function where outputs are distributed with Poisson distribution with given λ
- generated quantities block for generating values for predictions

Analysis of samples:







Data is consistent with prior predictive distribution. Most accidents were generated as expected for the number of 3 and 4 accidents.

Lambda distribution is much thinner comparing to the prior. The reason is wide half normal distribution that isn’t too informative.
There were not any issues with the sampling.

6. Prior second model

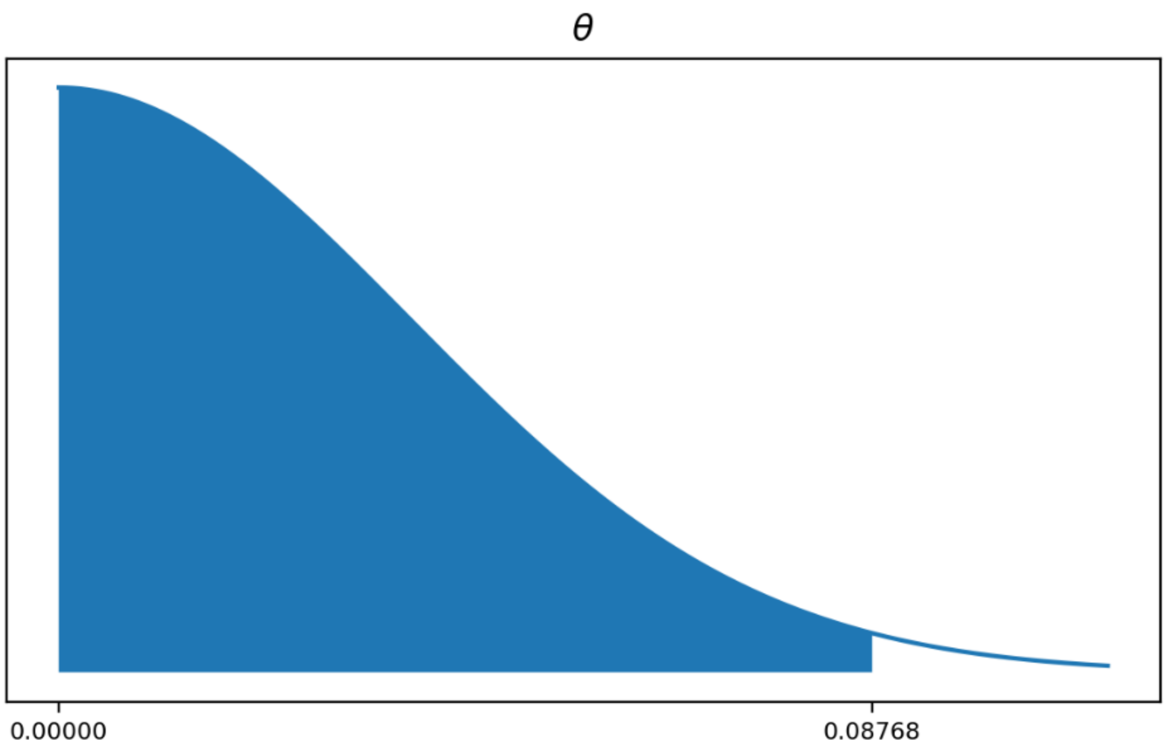
To compute the bound we will use previous argument but now λ will be adjusted

$$\lambda = \theta * \bar{n}$$

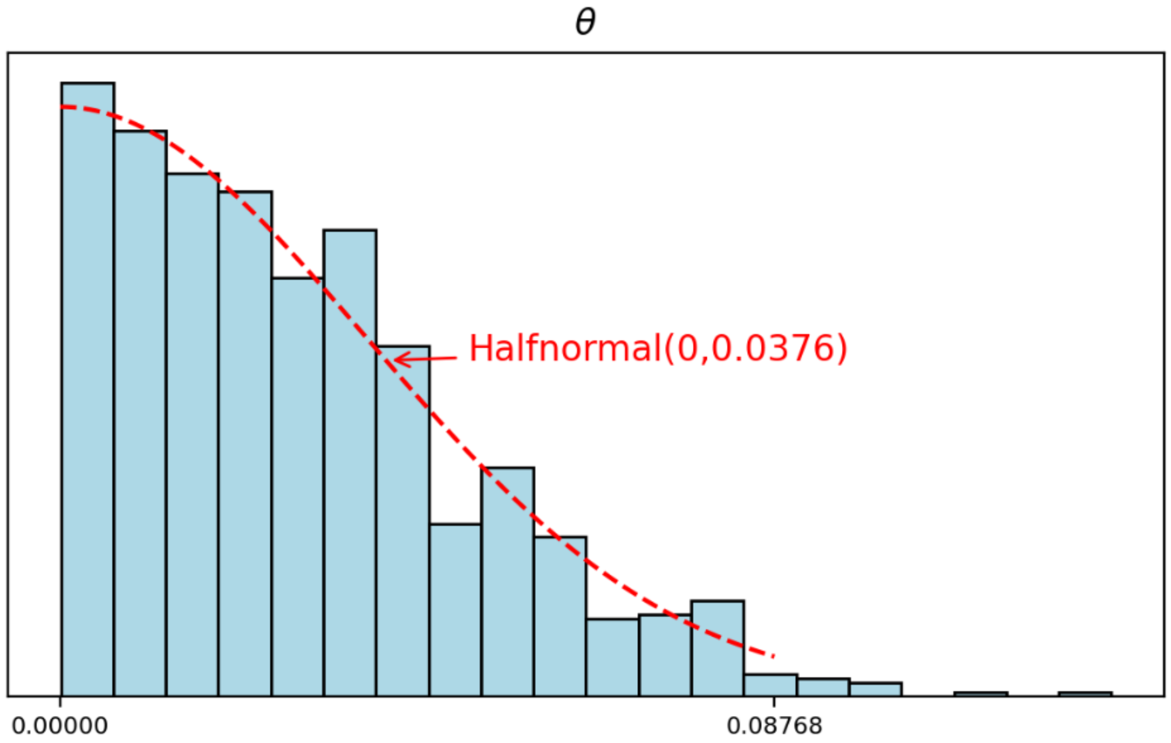
so the condition will look like this:

$$\theta * \bar{n} + 3 * \sqrt{\theta * \bar{n}} \approx 48$$

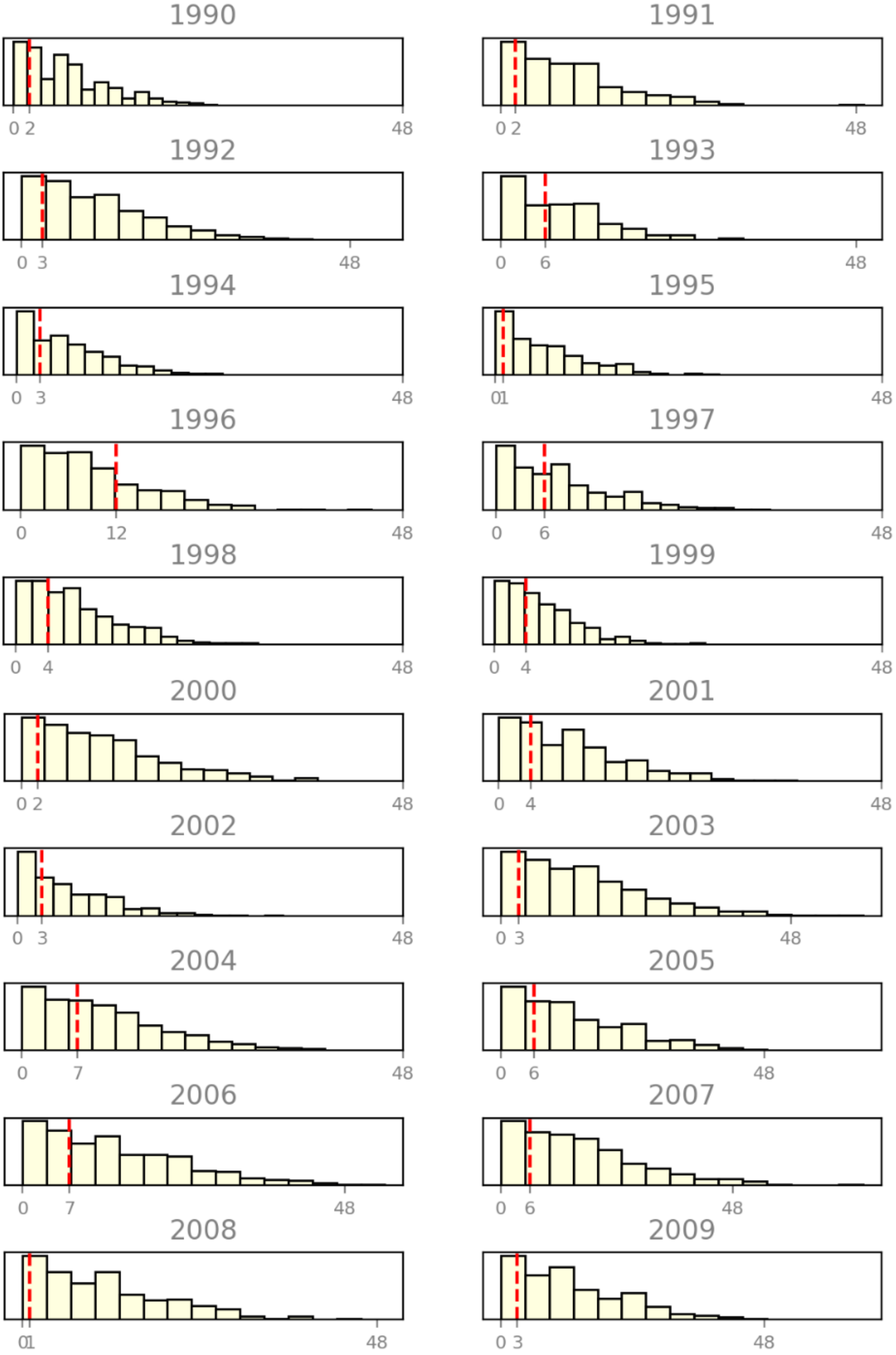
The standard deviation for such a distribution using the Newton's method is 0,0377.

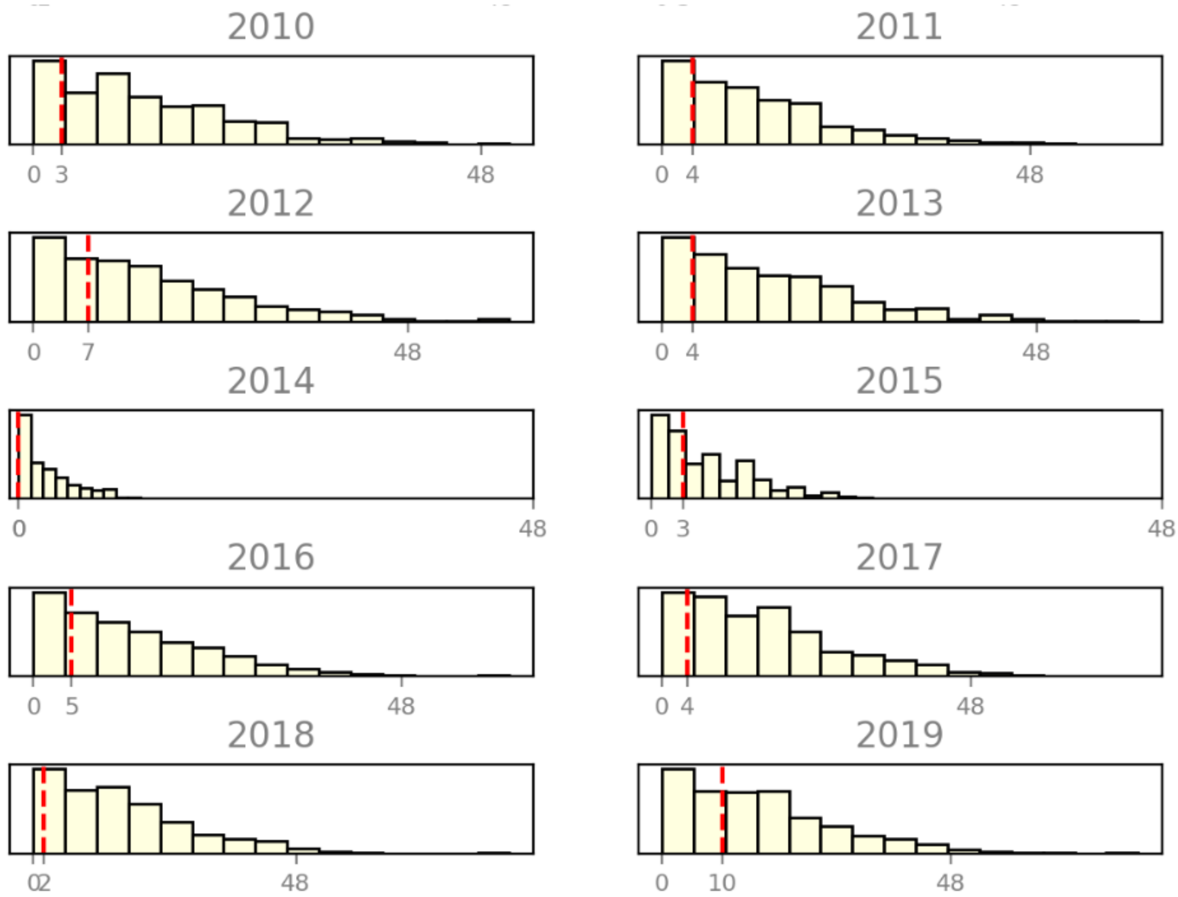


Prior predictive checks



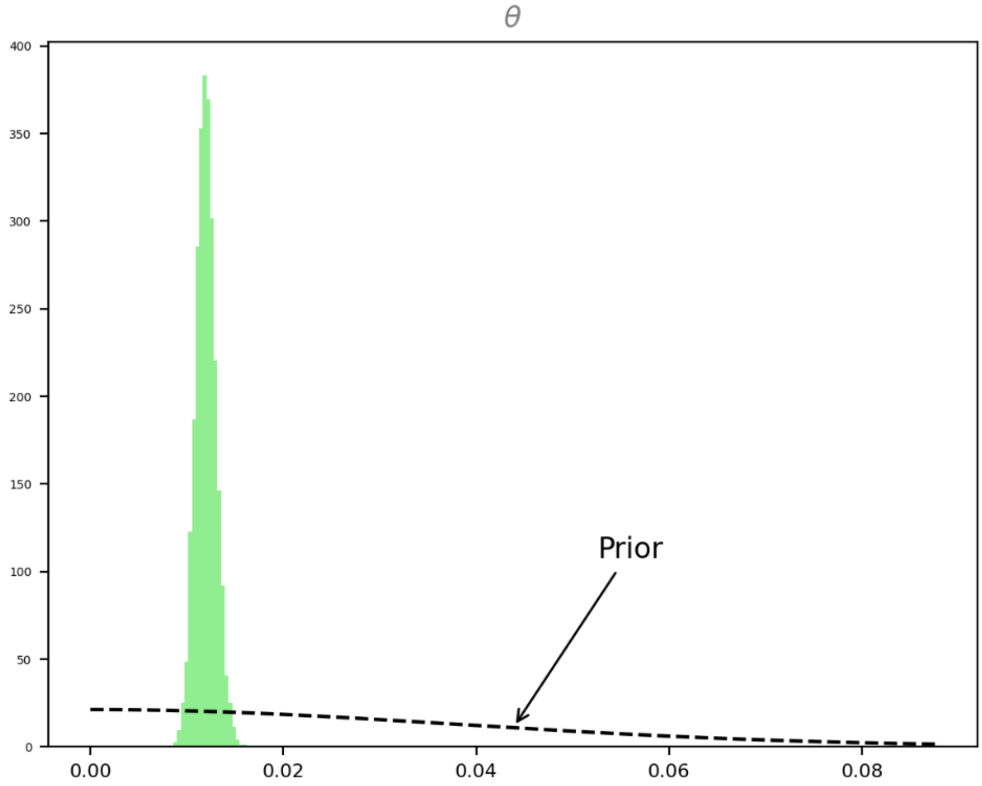
Prior distribution of deaths and actual value

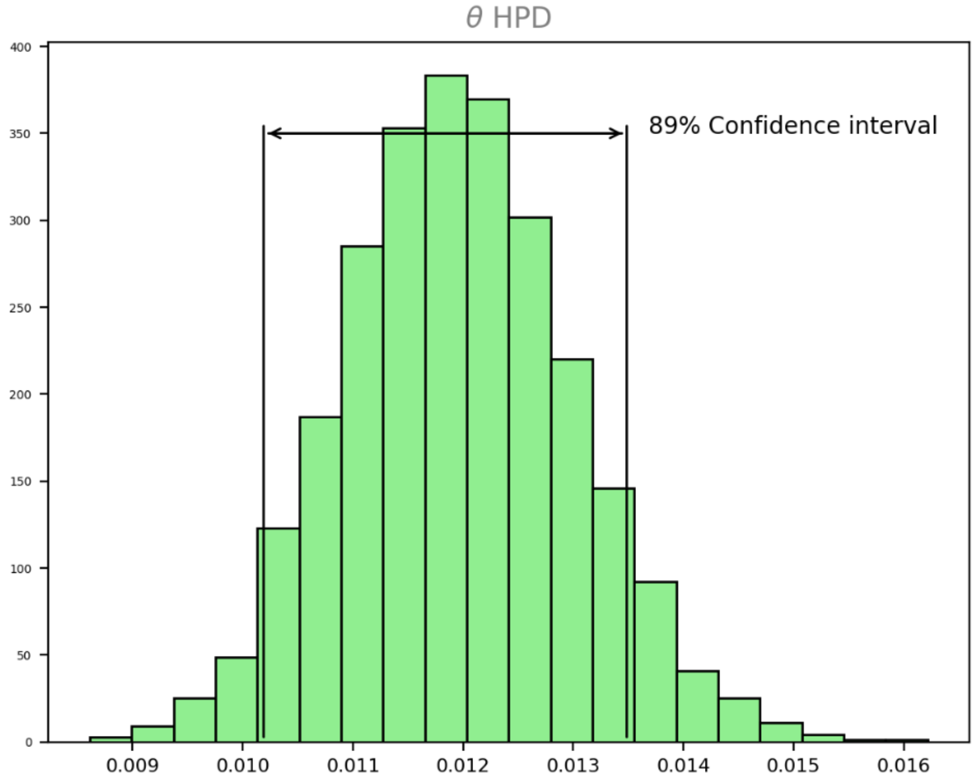




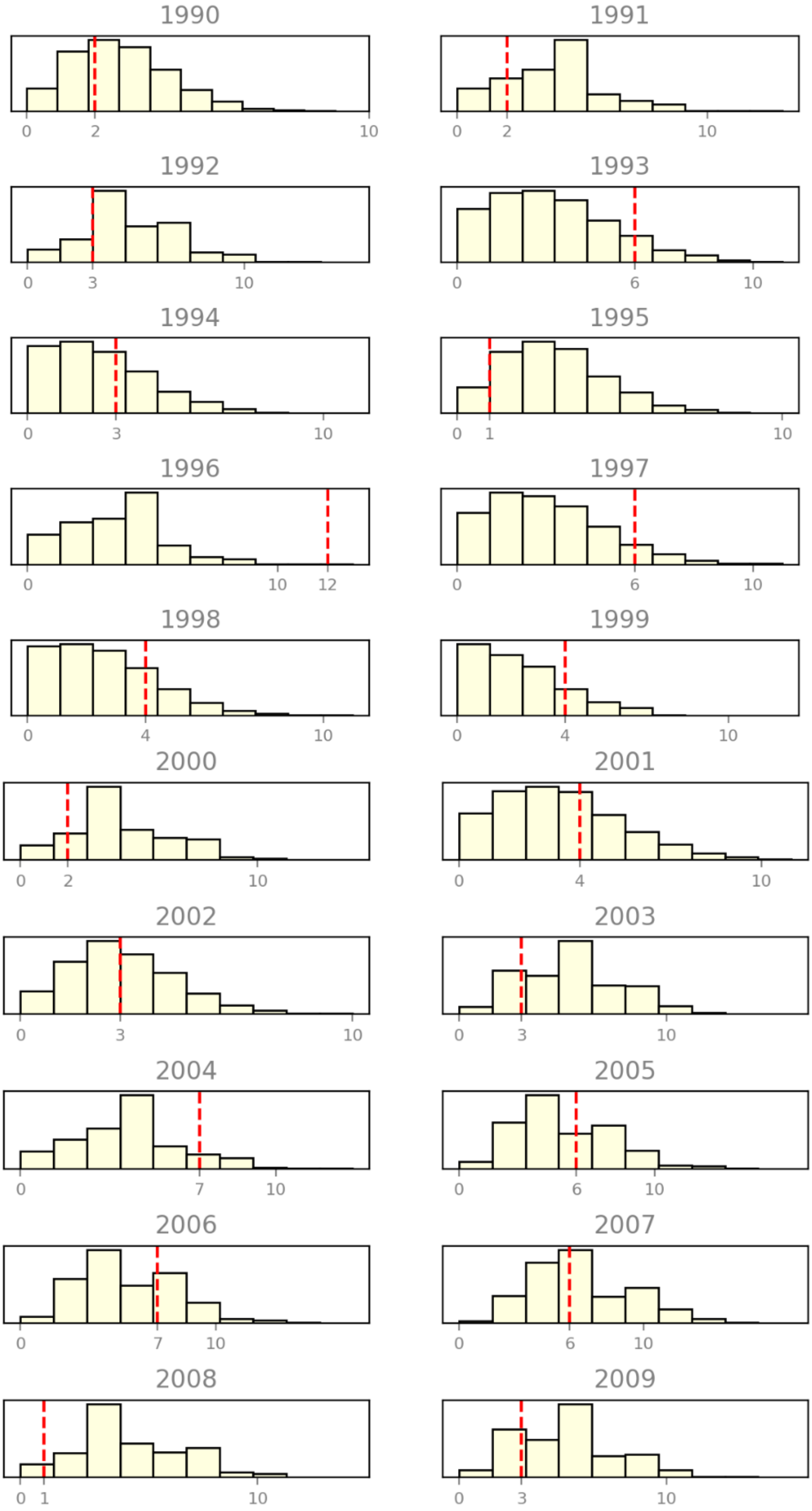
Generated number of deaths is located in the most probable area so prior is consistent with data.

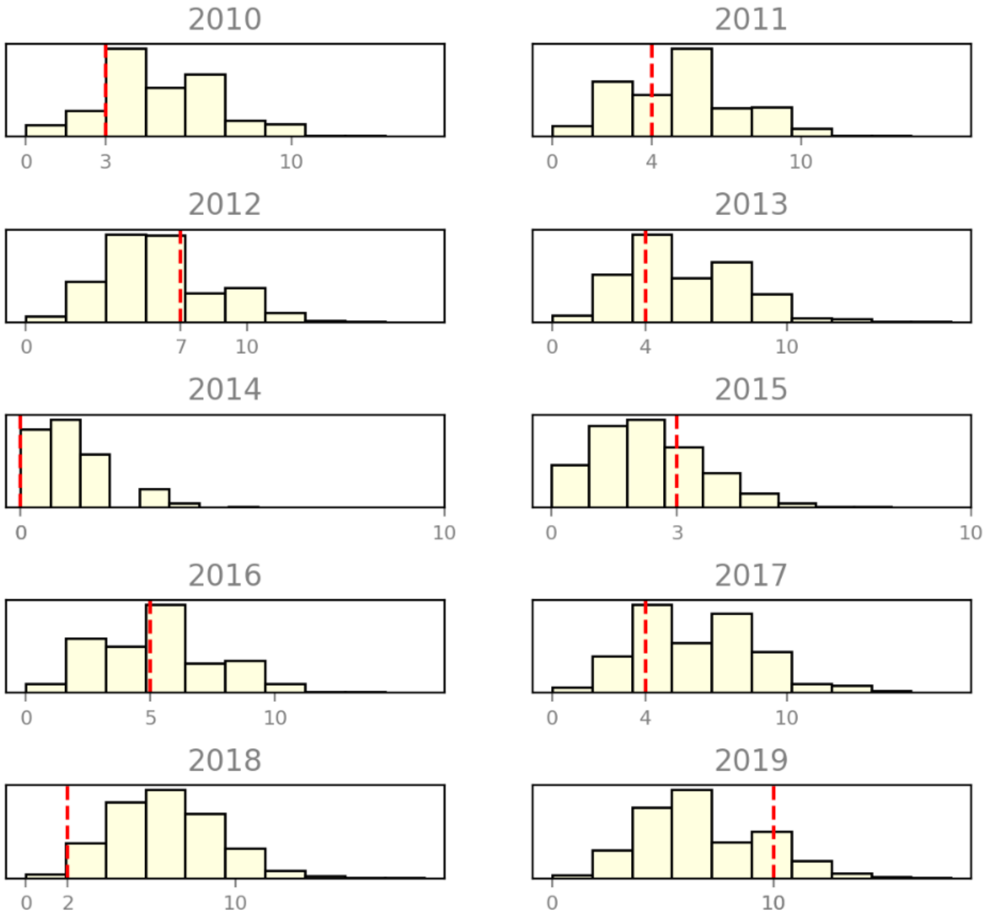
7. Posterior predictive checks





Posterior distribution of deaths and actual value





Data is consistent with the prior. Years with fewer climbers climbing tend to have a thinner distribution, with much of the probability density at low values, and for such cases, the number of accidents is low. Generated distribution for a given year is not consistent with actual data. The reason for such behavior may be additional conditions that have not been specified in the model. An example is 2014, where despite the prediction of which the 89% probability density interval does not include the actual value of the number of deaths. This is due to an avalanche whose occurrence is not modeled due to insufficient data.

There were not any issues with the sampling.

8. Model comparison

Leave-one-out cross-validation (LOO) and the widely applicable information criterion (WAIC) are methods used to estimate the out-of-sample prediction accuracy of a Bayesian model. They do this by evaluating the log-likelihood of the model at the posterior simulations of the parameter values.

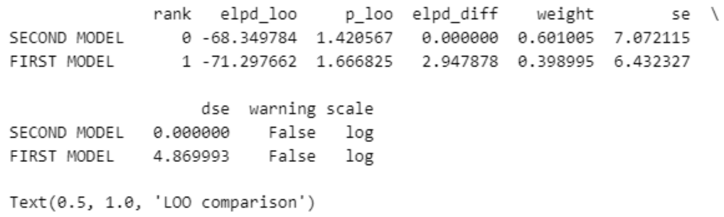
Both LOO and WAIC assess the quality of a statistical model based on two criteria:

- Model fit to the data: LOO and WAIC examine how well the model fits the available data by evaluating the log-likelihood based on the posterior simulations of the parameter values.

- Model complexity: LOO and WAIC also take into account the complexity of the model, including the number of parameters and the intricacy of the model structure, to assess how well the model generalizes to new data.

By considering both the fit to the data and the complexity of the model, LOO and WAIC provide information about the quality of the model in terms of predicting new data.

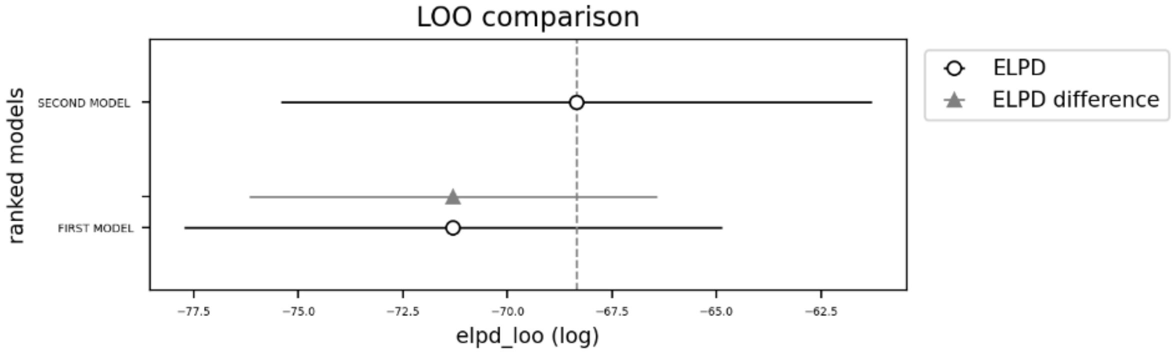
LOO comparision



```
              rank   elpd_loo     p_loo  elpd_diff    weight        se  \
SECOND MODEL     0 -68.349784  1.420567   0.000000  0.601005  7.072115
FIRST MODEL      1 -71.297662  1.666825   2.947878  0.398995  6.432327

                   dse  warning scale
SECOND MODEL  0.000000    False   log
FIRST MODEL   4.869993    False   log

Text(0.5, 1.0, 'LOO comparison')
```

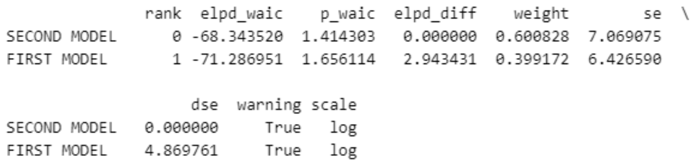


Based on the LOO criteria - taking into consideration:

- by the rank - second model is better (lower score)
- by the probability of correctness - second model is better (higher weight equal to 0,601005)

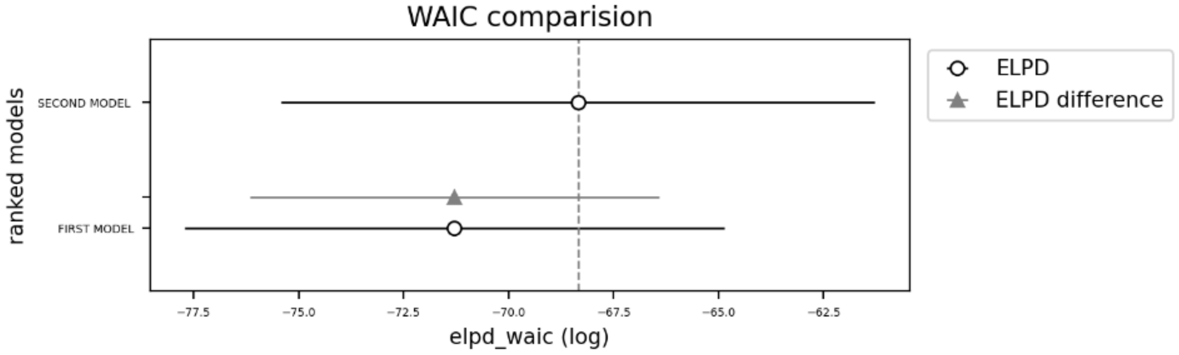
- by the out-of-sample predictive fit - first model is better (higher p_loo value)
- by the standard error of the difference information criteria - small difference between models (se and dse value)

WAIC comparision



```
              rank  elpd_waic    p_waic  elpd_diff    weight        se  \
SECOND MODEL     0 -68.343520  1.414303   0.000000  0.600828  7.069075   
FIRST MODEL      1 -71.286951  1.656114   2.943431  0.399172  6.426590   

                   dse  warning scale  
SECOND MODEL  0.000000     True   log  
FIRST MODEL   4.869761     True   log  
```



Results from WAIC comparisons are very similar, values are almost identical.

9. Conclusions

Visually based on histograms it is hard to assess which model is better. On the other hand, basing on the WAIC and LOO comparison we can conclude that the second model is better. We agree with this outcome, because second model uses more data.

Predicting fatal accidents on Mount Everest is not an easy task for several reasons. Firstly, the mountain the highest peaks in the world, and extreme weather conditions such as strong winds, low temperatures, and sudden weather changes pose significant risks to climbers. Secondly, each Everest expedition is a unique challenge, and factors such as physical fitness, climbing experience, supplies, and medical conditions can all impact the safety of participants. Additionally, difficulties in accessing the mountain and the limited window of favorable weather for summit attempts make planning and predicting accidents more complex. All these factors make predicting accidents on Mount Everest a challenging task that requires considering multiple variables and is difficult to accurately estimate.