Bayesian Statistics
Introduction

Peter Büchel

HSLU W

SA: Week 10

Bayesian Statistics

- The focus of this module will be on Bayesian statistics

> *When the Facts Change, I Change My Mind. What Do You Do, Sir?*
>
> John Maynard Keynes, Economist

- Bayesian statistics: Unified approach to statistics
- Much more natural approach of statistics for machine learning
- We have a model, collect data and "learn" from it using so-called Bayesian inference
- Very intuitive and is the mathematical version of how we think (usually)

Example

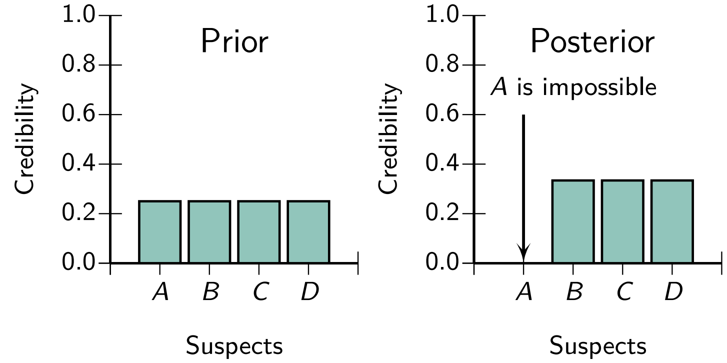
- Bayesian inference: *Reallocation* of credibility across possibilities
- Suppose we step outside one morning and notice that the sidewalk is wet, and wonder why
- We consider all possible causes of the wetness, including recent rain, recent garden irrigation, a newly erupted underground spring, a broken sewage pipe, a passerby who spilled a drink, and so on
- If all we know until this point is that some part of the sidewalk is wet, then all those possibilities will have some prior credibility based on previous knowledge
- For example, recent rain may have greater prior probability than a spilled drink from a passerby

- We look around and collect new information
- If we observe that the sidewalk is wet for as far as we can see, as are the trees and parked cars, then we re-allocate credibility to the hypothetical cause of recent rain
- The other possible causes, such as a passerby spilling a drink, would not account for the new information
- If instead we observed that the wetness was localised to a small area, and there was an empty drink cup a few feet away, then we would re-allocate credibility to the spilled-drink hypothesis, even though it had relatively low prior probability
- This sort of reallocation of credibility across possibilities is the essence of Bayesian inference

Example

- Suppose a murder has happened
- Four people A, B, C and D have made death threats against the murdered person
- *Assume* that one of the these suspect is the murderer and they didn't know each other, so it's not possible that, say, B *and* D committed the murder
- That is all we know
- We assign credibilities to all suspects of having committed the murder
- If credibility of a suspect is zero: Suspect definitely not the murderer
- If credibility of a suspect is one: Suspect definitely the murderer

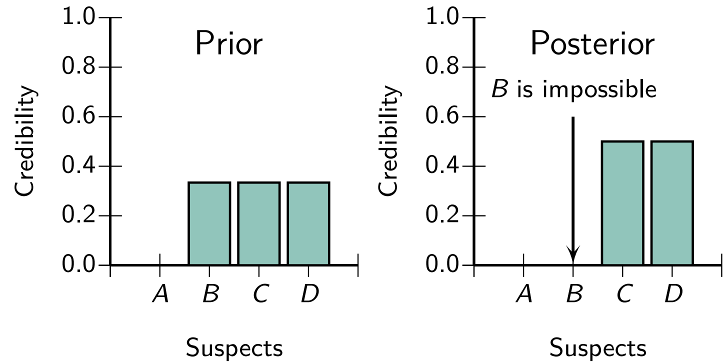
- Know nothing more about the suspects: Assign to each of them the credibility 0.25 of having committed the murder (see Figure left)



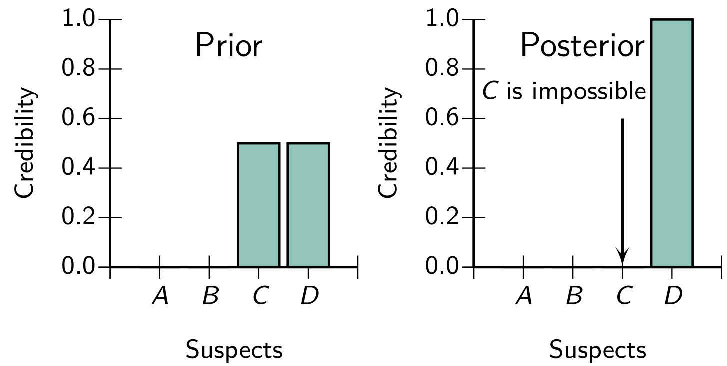
- We call this the *prior distribution*: Our knowledge about suspects *before* we are looking for evidence
- Now, we want to question suspect A, but realise that he has emigrated to a far away island

- The island administration confirms that he was on the island at the time of the murder
- Hence we can rule out suspect A as murderer
- Of course: Credibility of the other three suspects changes
- New credibilities of having committed the murder: 0.333 each (see Figure right)
- We call this the *posterior distribution*: Our knowledge about the suspects *after* we collected some data
- By collecting data, we have been able to reallocate the credibilities from the prior to the posterior distribution
- This idea lies at the heart of Bayesian inference

- We are not finished yet
- What about the other murder supects?
- Since suspect A is not a suspect anymore, we can concentrate on the other three suspects
- Use old posterior distribution as *new prior* distribution (Figure left)



- We read in the newspaper that suspect B has died in a car accident before the murder and the coroner confirms this
- Reallocate credibilities again: New posterior distribution
- Credibilities for the suspects C and D are now 0.5 each
- Repeat process again:
  - Old posterior becomes new prior distribution
  - Collect data
  - Reallocate the credibilities: New posterior distribution

- 

- Suspect C is in prison for an unrelated crime and the prison administration confirms this
- Therefore suspect D is the murderer
- Example completely idealised, but shows idea behind Bayesian inference quite nicely
- Data are noisy: Never have a completely deterministic process and hardly ever end up with certainty

Bayes' Theorem: Prior, Likelihood, Posterior

- Already seen the Bayes' theorem before
- Recapitulate the Bayes' theorem and introduce a new interpretation

Example

- Table: Proportions of combinations of hair colour and eye colour

	Hair color				
Eye color	Black	Brunette	Red	Blond	Marginal (eye color)
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	**0.14**	0.03	0.16	**0.36**
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.12	0.21	1

- The colored value 0.14 is probability

$$P(\text{Blue} \cap \text{Brunette})$$

- Probability person has blue eyes *and* brunette hair

- The value 0.36 at the end of the second row:

$$P(\text{Blue}) = 0.36$$

- Probability that a person has blue eyes
- Consider only the second row:

Eye color	Black	Hair color Brunette	Red	Blond	Marginal (eye color)
Blue	0.03/0.36	**0.14**/0.36	0.03/0.36	0.16/0.36	**0.36**/0.36 = 1

- Determine probability that a person with blue eyes has brunette hair
- For this probability: Divide values in second row by $P(\text{Blue}) = 0.36$

- Denote probability that person with blue eyes has brunette hair with

$$P(\text{Brunette} \mid \text{Blue})$$

- Therefore

$$P(\text{Brunette} \mid \text{Blue}) = \frac{P(\text{Blue} \cap \text{Brunette})}{P(\text{Blue})} = \frac{0.14}{0.36} = 0.39$$

- Hence 39 % of the people with blue eyes have brunette hair

- Add up all values in the second row of first table: $P(\text{Blue})$

$P(\text{Blue}) = P(\text{Blue} \cap \text{Black}) + P(\text{Blue} \cap \text{Brunette}) + P(\text{Blue} \cap \text{Red}) + P(\text{Blue} \cap \text{Blond})$

- With definition of conditional propbability, say, $P(\text{Blue} \cap \text{Brunette})$:

$$P(\text{Blue} \cap \text{Brunette}) = P(\text{Brunette} \mid \text{Blue}) \cdot P(\text{Blue})$$

- Rewrite $P(\text{Blue})$:

$$\begin{aligned} P(\text{Blue}) = {} & P(\text{Black} \mid \text{Blue}) \cdot P(\text{Blue}) + P(\text{Brunette} \mid \text{Blue}) \cdot P(\text{Blue}) \\ & + P(\text{Red} \mid \text{Blue}) \cdot P(\text{Blue}) + P(\text{Blond} \mid \text{Blue}) \cdot P(\text{Blue}) \end{aligned}$$

- Call the probability $P(\text{Blue})$ a *marginal* probability

In general

- Table:

Row	$\cdots$	Column c	$\cdots$	Marginal
$\vdots$		$\vdots$		
r	$\cdots$	$P(r \cap c) = P(r \mid c)P(c)$	$\cdots$	$P(r) = \sum_{c^*} P(r \mid c^*)P(c^*)$
$\vdots$		$\vdots$		
Marginal		$P(c)$		

- In general:

$$P(r \cap c) = P(r \mid c)P(c)$$

- But also:

$$P(r \cap c) = P(c \mid r)P(r)$$

- Therefore:

$$P(c \mid r)P(r) = P(r \mid c)P(c)$$

- Hence:

$$P(c \mid r) = \frac{P(r \mid c)P(c)}{P(r)}$$

- This is Bayes' theorem
- Describes $P(c \mid r)$ in terms of $P(r \mid c)$, $P(c)$ and $P(r)$
- Rewrite the marginal probability $P(r)$ for n columns

$$P(r) = P(r \mid c_1)P(c_1) + P(r \mid c_2)P(c_2) + \ldots + P(r \mid c_n)P(c_n)$$

- Pluggin the last expression into the Bayes' theorem we obtain the Bayes' theorem the way we will use it.

Bayes' theorem

$$P(c \mid r) = \frac{P(r \mid c)P(c)}{P(r \mid c_1)P(c_1) + P(r \mid c_2)P(c_2) + \ldots + P(r \mid c_n)P(c_n)}$$

Likelihood, prior, posterior

- Introduction how to learn from data or information using Bayes' theorem

Example

- Malaria is a life-threatening disease
- It is characterised, among other things, by high, periodically occurring fever attacks in affected people
- Doctors measure parameters such as body temperature, blood counts, cardiograms or use medical tests
- OptiMAL is such a test to quickly detect malaria

- It cannot guarantee that it will be positive if a person has malaria
- Clinical trials have shown that the OptiMAL test has a probability of 0.917 (or 91.7 %) of being positive in people infected with malaria
- This probability: Called the *sensitivity* or *true positive rate* of the test
- It is therefore the probability that the test will be positive, given the person has malaria
- This can be written as conditional probability:

$$P(+ \mid M) = 0.917$$

- Similarly, a medical test should react as negatively as possible if the person examined is not affected by the disease
- How well a test does this is indicated by the *specificity* (also *true negative rate*)
- Probability that a test will be negative, given the person is not ill
- In the OptiMAL test, the specificity (from clinical trials):

$$P(- \mid \overline{M}) = 0.935$$

- A doctor can use the OptiMAL test on a person
- Test is positive
- What is the probability that the patient has malaria?

- So what we are looking for is the conditional probability:

$$P(M \mid +) = ?$$

- According to Bayes' theorem:

$$P(M \mid +) = \frac{P(+ \mid M) \cdot P(M)}{P(+)} = \frac{P(+ \mid M) \cdot P(M)}{P(+ \mid M) \cdot P(M) + P(+ \mid \overline{M}) \cdot P(\overline{M})}$$

- To determine the probability, need $P(M)$, the probability that this person has malaria
- We don't know anything about the person and visit WHO's website
- About 3 %:

$$P(M) = 0.03$$

- Hence:

$$\begin{aligned} P(M \mid +) &= \frac{P(+ \mid M) \cdot P(M)}{P(+ \mid M) \cdot P(M) + P(+ \mid \overline{M}) \cdot P(\overline{M})} \\ &= \frac{0.917 \cdot 0.03}{0.917 \cdot 0.03 + 0.065 \cdot 0.97} \\ &= 0.303\,776\,5 \end{aligned}$$

- *Before* the test was taken, $P(M)$ for that person was 0.03
- *After* the test is positive (i.e., got additional information) the probability that this person has malaria is about 0.3 or 30 %
- Introduce the following terms:
  - $P(M)$: *Prior* probability
  - $P(M \mid +)$: *Posterior* probability
  - $P(+ \mid M)$: *Likelihood* function

- Test is not particularly convincing, but additional information has resulted in a tenfold increase of probability that this person has malaria
- However, person is now still unsure and wants to know more precisely whether she has malaria or not
- Does the same test again (not recommended, see DoE)
- However: $P(M)$ is no longer 0.03 but 0.304, since test has changed the probability
- Choose posterior-probability as new prior-probability
- All other variables remain the same (at least we assume so)

- The person receives another positive test:

$$
\begin{aligned}
P(M \mid +) &= \frac{P(+ \mid M) \cdot P(M)}{P(+ \mid M) \cdot P(M) + P(+ \mid \overline{M}) \cdot P(\overline{M})} \\
&= \frac{0.917 \cdot 0.304}{0.917 \cdot 0.304 + 0.065 \cdot 0.696} \\
&= 0.86
\end{aligned}
$$

- This person now has malaria with a probability of 0.86
- Of course, the person can take the test again
- In this case, we choose 0.86 as the new prior probability
- If test is positive again, the probability that this person has malaria increases to 0.99
- Seen in this example how new data or information (positive tests) using Bayes theorem results in an adjustment of one parameter - namely the probability of having malaria

Coin tosses: Who cares?

- Following examples: Introduce mathematics in more detail
- Example: Coin tosses
- The fairness of a coin may be of great importance in high stakes games, but unimportant else
- So why bother studying the statistics of coin tosses?

- Because coin tosses are a substitute for countless other events in real life that matter to us
  - For a certain type of heart surgery, we can classify the patient's outcome as "survived" or "not survived" depending on whether or not they survived more than a year. We want to know how likely patients are to survive more than one year
  - For a given type of drug, we can describe the outcome as "headache" or "no headache", where we want to quantify the probability of headache
  - In a survey, the result could be "agree" or "disagree", and we want to know the probability of each answer
  - What is the probability of two planes crashing in midair?
- Basically whenever there is a yes/no decision to a question, coin tosses can be used as a model

Example: Coin tosses

- Probability of flipping "head" H: θ
- Probability of flipping "tail" T: $1-\theta$
- Consider a new one Swiss franc coin from the Swiss National Bank
- The question is now, how large is θ?
- Because the coin comes from the reputable National Bank, is new and looks symmetrical, we may assume that the coin is *fair*, i.e., $\theta = 0.5$
- Now a *real* coin is never fair in the sense that θ is *exactly* 0.5, namely $\theta = 0.500\,000\,0\ldots$

- It may be:
$$0.48 < \theta < 0.52$$
- Note that there are infinitely many values for θ, ranging from 0 to 1
- Want to make a statement based on data about how large θ actually is
- We now discuss two ways of doing this:
  - Either *frequentistic* (later): Flip the coin a *lot* of times and note how many times H has been flipped
  - Using *Bayes' theorem*

- Flip the coin *once* and get H: Denote by $y = 1$
- Collected data about the coin and want to know about θ
- Are looking for the conditional probability:
$$P(\theta \mid y = 1)$$
- Apply Bayes' theorem:
$$P(\theta \mid y = 1) = \frac{P(y = 1 \mid \theta) \cdot P(\theta)}{P(y = 1)}$$
- On the right hand: Three unknown quantities:
$$P(y = 1 \mid \theta), \qquad P(\theta) \qquad \text{und} \qquad P(y = 1)$$
- Call:
  - $P(\theta)$: Prior probability
  - $P(y = 1 \mid \theta)$: Likelihood function
  - $P(y = 1)$: Margin probaility

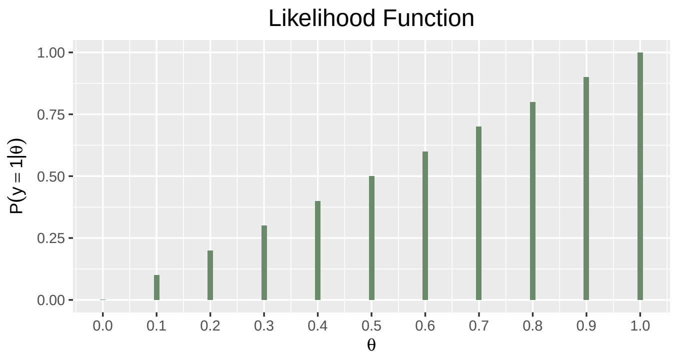
- Although θ can take infinitely many values, discretise θ for now
- Assume that θ can only take values of $0, 0.1, 0.2, \ldots, 0.9, 1$
- Likelihood function $P(y = 1 \mid \theta)$ depends on θ:
$$P(y = 1 \mid \theta) = \theta$$
- If the probability of tossing H is θ (first θ), then with a probability of θ (second θ) H also appears
- For example, if $\theta = 0.8$, then
$$P(y = 1 \mid 0.8) = 0.8$$
- Probability of tossing head given probability of 0.8 is, of course, also 0.8

- For multiple θ's, there are also multiple $P(y = 1 \mid \theta)$'s
- The likelihood values for $P(y = 1 \mid \theta)$:

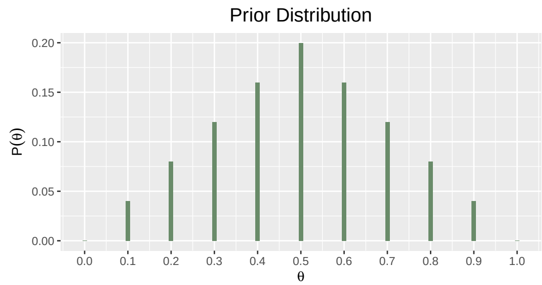
```
like <- seq(0, 1, 0.1)
like

[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

- Sketch:



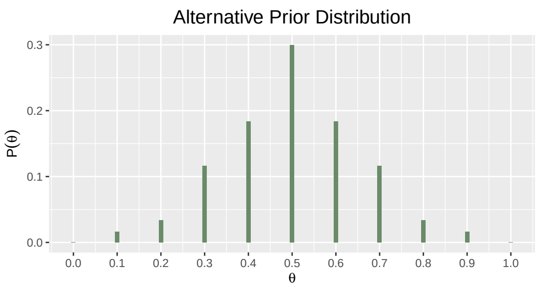
- *Define* prior distribution $P(\theta)$, and there are many possibilities
- Have to define it *before* we collect data
- If we think of the coin as rather fair, but do not know for sure, we can *choose* a prior distribution:



- The prior values for the θ's are:

```
y <- c(0, 1, 2, 3, 4, 5, 4, 3, 2, 1, 0)
prior <- y / sum(y)
prior
[1] 0.00 0.04 0.08 0.12 0.16 0.20 0.16 0.12 0.08 0.04 0.00
```


Remarks

- For $\theta = 0$ and $\theta = 1$: Probabilities are 0 in prior distribution, since we want to exclude that only H or only T can occur
- 

  Why not choose the following prior distribution?
- Later in more detail: The effect of different prior distributions on posterior distribution

- Calculate the marginal probability $P(y = 1)$:

$$
\begin{aligned}
P(y=1) &= P(y=1 \mid \theta=0) \cdot P(\theta=0) + \ldots + P(y=1 \mid \theta=1) \cdot P(\theta=1) \\
&= \sum_{i=0}^{10} P(y=1 \mid \theta_i) \cdot P(\theta_i)
\end{aligned}
$$

- **R**:

```
margin <- sum(prior * like)
margin

[1] 0.5
```

- Posterior probabilities for *all* θ's, i.e., the posterior distribution
- For example:

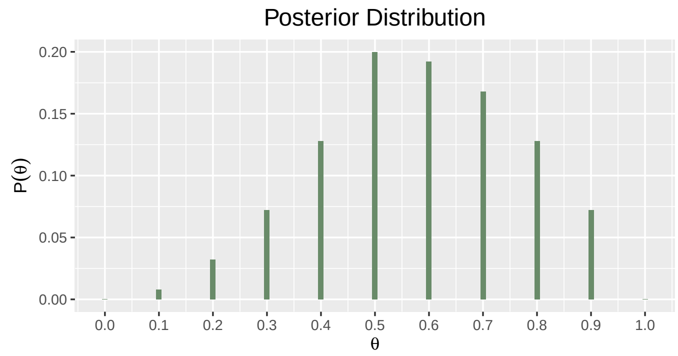
$$P(0.8 \mid y=1) = \frac{P(y=1 \mid 0.8) \cdot P(0.8)}{P(y=1)} = \frac{0.8 \cdot 0.08}{0.5} = 0.128$$

- R:

```
post = prior * like / margin
post

[1] 0.000 0.008 0.032 0.072 0.128 0.200 0.192 0.168 0.128 0.072 0.000
```


- Sketch:



- Prior distribution influences the posterior distribution “by a slightly different weighting” compared to the prior distribution
- The θ's above 0.5 are weighted more heavily than those below 0.5
- Tossing H: Data gives indication for higher probabilities of tossing H
- Tossing T: Higher weighting for tossing probabilities below 0.5

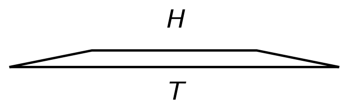
Choice of prior distribution

- One or *the* critical point in example: Choice of prior distribution
- Seems to be something subjective, since we can *choose* it
- The (seemingly arbitrary) choice of prior distribution used to (and still does) lead to huge debates about whether Bayes statistics is applicable at all
- Well, success proves the Bayes approach right

- Although there is something subjective about the choice of the prior distribution, the choice is *not arbitrary*
- Different people will choose the prior distribution differently for a given problem
- However, this distribution should be similar
- Above all, the choice of prior distribution should be reasonable

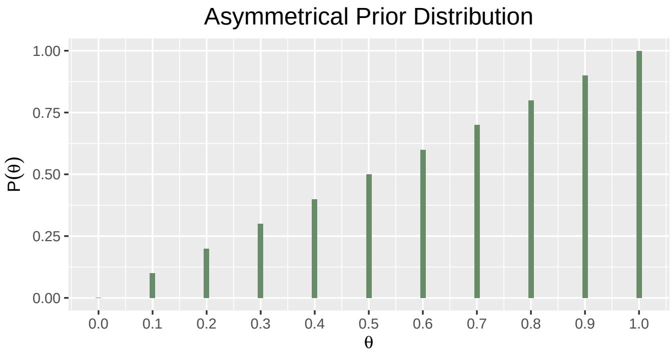
Example

- Consider a coin that has an asymmetric cross-section:



- A symmetric prior distribution makes no sense because we already know beforehand that this kind of distribution cannot be correct
- In this case: Much more often T than H

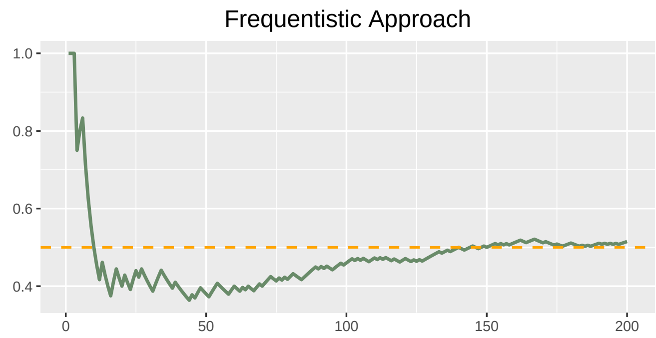
- Choose a prior distribution that could be similar to:



- Choice of prior: Put into the prior distribution all available information
- This information will be uncertain, but it is the best we have
- Then apply Bayes' theorem

Frequentistic approach

- Frequentist approach: Repeat an experiment virtually endlessly



- Dividing the number of H observed by the number of trials
- Expectation: Value increasingly approximates true probability θ
- Here: $\theta = 0.5$

- Realistic for coins to have a very large number of trials
- But if coin is a model: In practice often not possible

Example

- You are developing a measuring device and want to determine its lifetime
- Ideally, the device should have a lifetime of, let's say, over 10 years
- Now, you want to sell the device today and not wait 10 years to see if the device will last that long
- The frequentist approach would work here, but it is not practical

Example

- As civil engineers, you cannot build 100 000 bridges to check how many have collapsed after 50 years
- The frequentist approach is of no use here, but the Bayesian approach is helpful

Example

- The history and future of the universe may be taken as a unique experiment - parallel universes elude our observation
- Bayes statistics is particularly relevant in cosmology for this reason

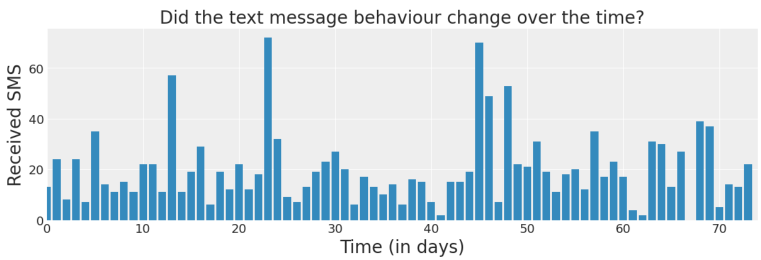
Example

- In the early days of aviation around 1955, an insurance company in the USA had wondered what the probability was that there would be a collision of two planes in the air
- Since there had been no serious collision of this kind up to that time, a prediction had to be made that was based on no experience
- Frequentistically: Probability was 0 at the time, simply because there had never been such a collision
- A statistician, L. H. Longley-Cook, took on this task

- Based on near misses and Bayes statistics, he concluded that despite the industry's safety record, there will be "between 0 and 4 collisions between aircraft in the next ten years"
- In sum, the company should prepare for a costly catastrophe by raising airline premium rates and buying reinsurance
- Two years later, his prediction proved correct
- A DC-7 and a Constellation collided over the Grand Canyon, killing 128 people in what was then the worst accident in commercial aviation
- Four years later, a DC-8 jet and a Constellation collided over New York City, killing 133 people in the planes and in the flats below

Example

- You are given a series of daily text-message counts from a user of your system
- Data, plotted over time



- You are curious to know if the user's text-messaging habits have changed over time, either gradually or suddenly
- How can you model this? (later)

- What are the prior's? (later)
- Posterior: