# Experience vs Data

Markus Gesmann
Düsseldorf Data Science Meetup, 13 June 2017

# My story for today

- How to asses risk with small data sets
- Three examples from insurance pricing
  - Using Bayes, Belief Networks and MCMC
  - R packages: gRain, RStan

# The insurance data conundrum

- Insurance companies have many customers
- But most customers have very few claims
- How do you price risk?

# How do you set the price?

- Three options:

  1. Start with the costs of the insured

  2. Start with the perceived value to the insured

  3. Call your competitor and ask for the price

# Example: Motor insurance

- We have clustered policyholders into ‘good’ and ‘bad’ drivers.

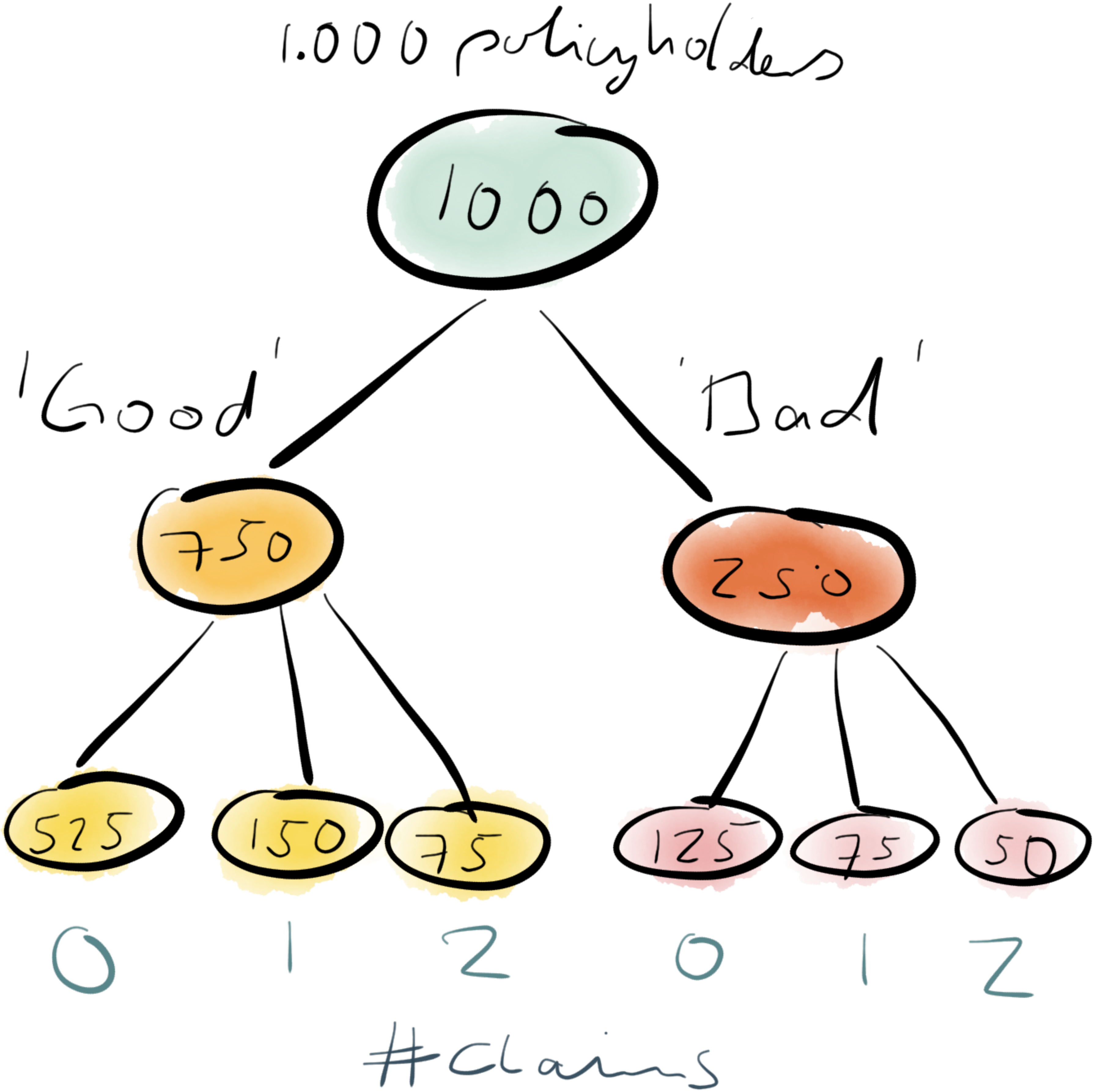
| Average number of claims per year | Frequency for ‘Good’ drivers | Frequency for ‘Bad’ drivers |
|---|---|---|
| 0 | 70% | 50% |
| 1 | 20% | 30% |
| 2 | 10% | 20% |

- Of our policyholders 75% are categorised as ‘good’, 25% as ‘bad’.

Source: Loss Models: From Data to Decisions

# Discussion

- How many claims would you expect from 1,000 policyholders in a year?
- How many claims would you expect from a random policyholder in a year?


1.000 policyholders
1000
'Good'
'Bad'
750
250
525
150
75
125
75
50
0
1
2
0
1
2
#claims

# Expected number of claims for one random policyholder

$$\begin{aligned} &75\%(0\cdot 70\% + 1\cdot 20\% + 2\cdot 10\%)+\\ &25\%(0\cdot 50\% + 1\cdot 30\% + 2\cdot 20\%)\\ =&0.475 \end{aligned}$$

# Customer asks for his renewal

- The customer is a policyholder of yours for the last two years.
- He had one claim over those two years.
- How many claims should we expect next year?

# Thomas Bayes can help

$$P(H|D) = \frac{P(H)\,P(D|H)}{P(D)}$$

What is our hypothesis?
What is our data?

H = “Customer is a ‘good’ driver”

D = “1 claim in two years”
= {(no claim in year 1 & one claim in year 2),
(one claim in year 1 & no claim in year 2)}
= {(1,0), (0,1)}

# Prior probability

$P(H) = 75\%$

# Likelihood

$$
\begin{aligned}
p(D \mid H) &= p(\{(1,0),(0,1)\} \mid H) \\
&= p(\{0\} \mid H)\, p(\{1\} \mid H) + \\
&\quad p(\{1\} \mid H)\, p(\{0\} \mid H) \\
&= 70\% \cdot 20\% + 20\% \cdot 70\% \\
&= 28\%
\end{aligned}
$$

# Data: Sum over all hypothesises

$$p(D) = \sum_i p(D | H_i)\, p(H_i)$$

# Data: Sum over all hypothesises

$$p(D) = p(D|H)p(H) + p(D|\overline{H})p(\overline{H})$$

$$= 28\% \cdot 75\% + (50\% \cdot 30\% + 30\% \cdot 50\%)25\%$$

$$= 28.5\%$$

# Probability that customer is a ‘good’ driver given that he had one claim in two years.

$$p(H|D) = \frac{p(H)\,p(D|H)}{p(D)}$$

$$= \frac{75\% \; 28\%}{28.5}$$

$$= 73.7\%$$

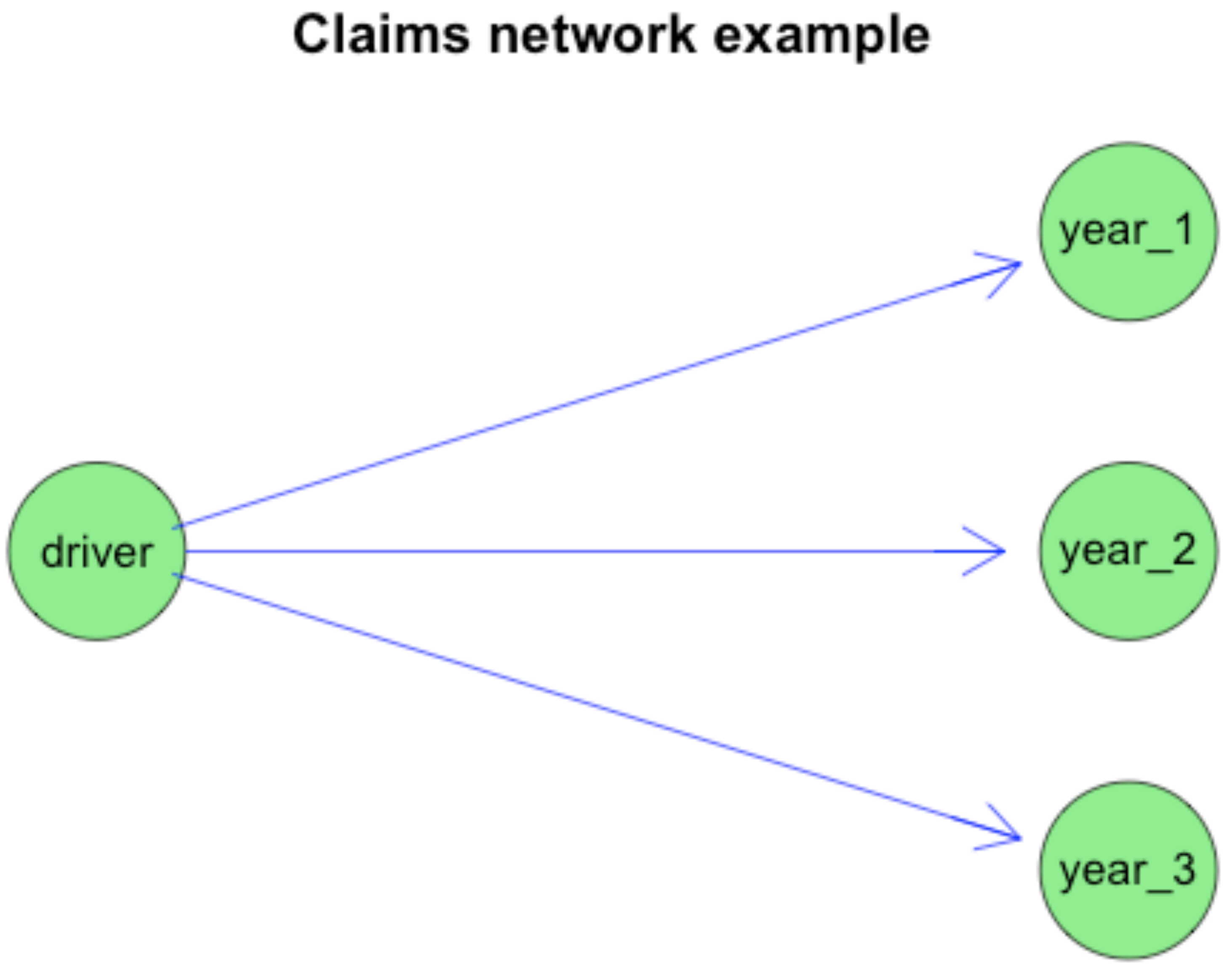
# Discussion

- What is the expected number of claims for this customer?

# Discussion

- Would you change your view of the customer, if you would know that he had the claim in year 1, but  not in year 2?
- Would you change your view, if the claim was in year 2, but not in year 1?

# Alternatively think of this as a belief network



# Define network in R

```
library(gRain)
library(Rgraphviz)
# Distribution of good and bad drivers
d <- cptable(~ driver, values=c(0.75, 0.25),
               levels=c("good","bad"))
claims <- c("0", "1", "2")
cond.prop <- c(0.7, 0.2, 0.1, 0.5, 0.3, 0.2)
c1 <- cptable(~ year_1|driver, values=cond.prop, levels=claims)
c2 <- cptable(~ year_2|driver, values=cond.prop, levels=claims)
c3 <- cptable(~ year_3|driver, values=cond.prop, levels=claims)
plist <- compileCPT(list(d, c1, c2, c3))
pn <- grain(plist)
plot(pn[["dag"]], main="Claims network example",
     attrs = list(node = list(fillcolor = "lightgreen"),
                  edge = list(color = "blue"),
                  graph = list(rankdir = "LR")))
```

# Set evidence

```
pn <- setEvidence(pn, nslist = list(year_1 = "0",
                              year_2 = “1"))

querygrain(pn, nodes = "driver", type = “marginal")
$driver
driver
    good      bad
0.7368421 0.2631579
```

# Hit and Run. What do you think?

- A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city.

- 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue.

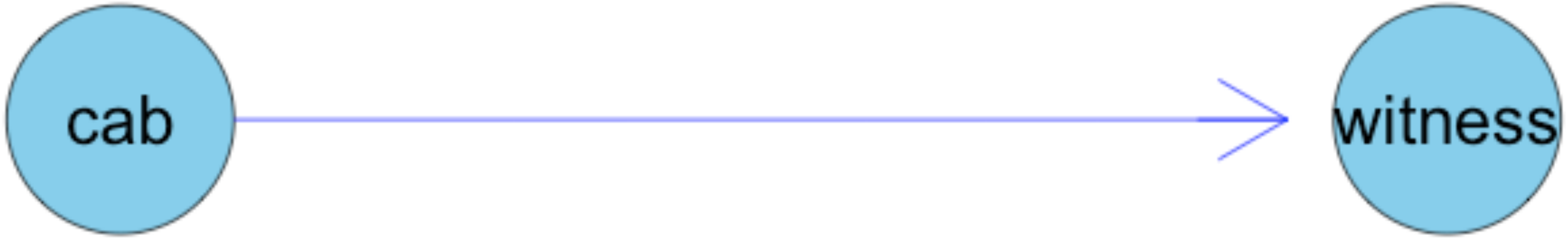
- The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colours 80% of the time and failed 20% of the time.

- What is the probability that the cab involved in the accident was Blue rather than Green knowing that this witness identified it as Blue?

Source: Daniel Kahneman: Thinking, Fast and Slow

H = Accident caused by Blue cab
D = Witness said the cab was Blue

## Thinking, fast and slow



# Use of belief network

```
cb <- cptable(~ cab, values=c(0.15, 0.85),
              levels=c("blue", "green"))
wtnss <- cptable(~ witness|cab,
                           values=c(0.8*0.15, 0.2*0.85,
                           0.8*0.85, 0.2*0.15),
                   levels=c("correct", "incorrect"))
plist <- compileCPT(list(cb, wtnss))
plist$witness
#            cab
# witness          blue      green
#    correct   0.4137931 0.95774648
#    incorrect 0.5862069 0.04225352
```

# Chances of entering a new product line successfully

- A company is planning to enter a new product line, where historically only 15% of their peers met their planned profits and 85% failed.

- The company has a track record of meetings its business plan profit targets 4 out 5 years.

- How much confidence would you have that this company can achieve its planned profit in the new product line?

# Predicting mid-air collisions

- The airline industry grew rapidly in the 1950s
- L.H. Longley-Cook was asked to price the risk for a mid-air collision of two planes
- All Longely-Cook knew was that there were no collisions in the previous 5 years

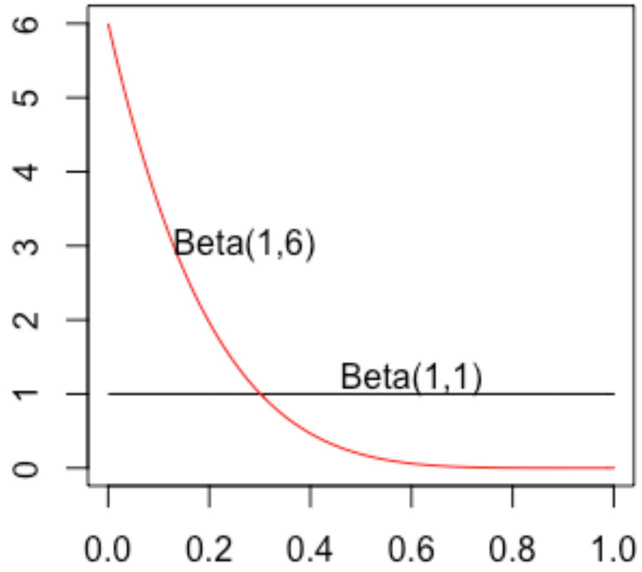
Source: Computational Actuarial Science with R

How do you think about this?

- Let's think of the years as a series of Bernoulli trials with unknown probability *p*

- That's a likelihood.

- Start with an uninformed prior, such as a Beta(α,β) with α=1, β=1 and mean $p_0 = \alpha/(\alpha + \beta)$=1/2



- Use the concept of conjugate prior to update my believe: $\alpha' = \alpha + \sum x_i = 1$, $\beta' = \beta + n - \sum x_i = 6$

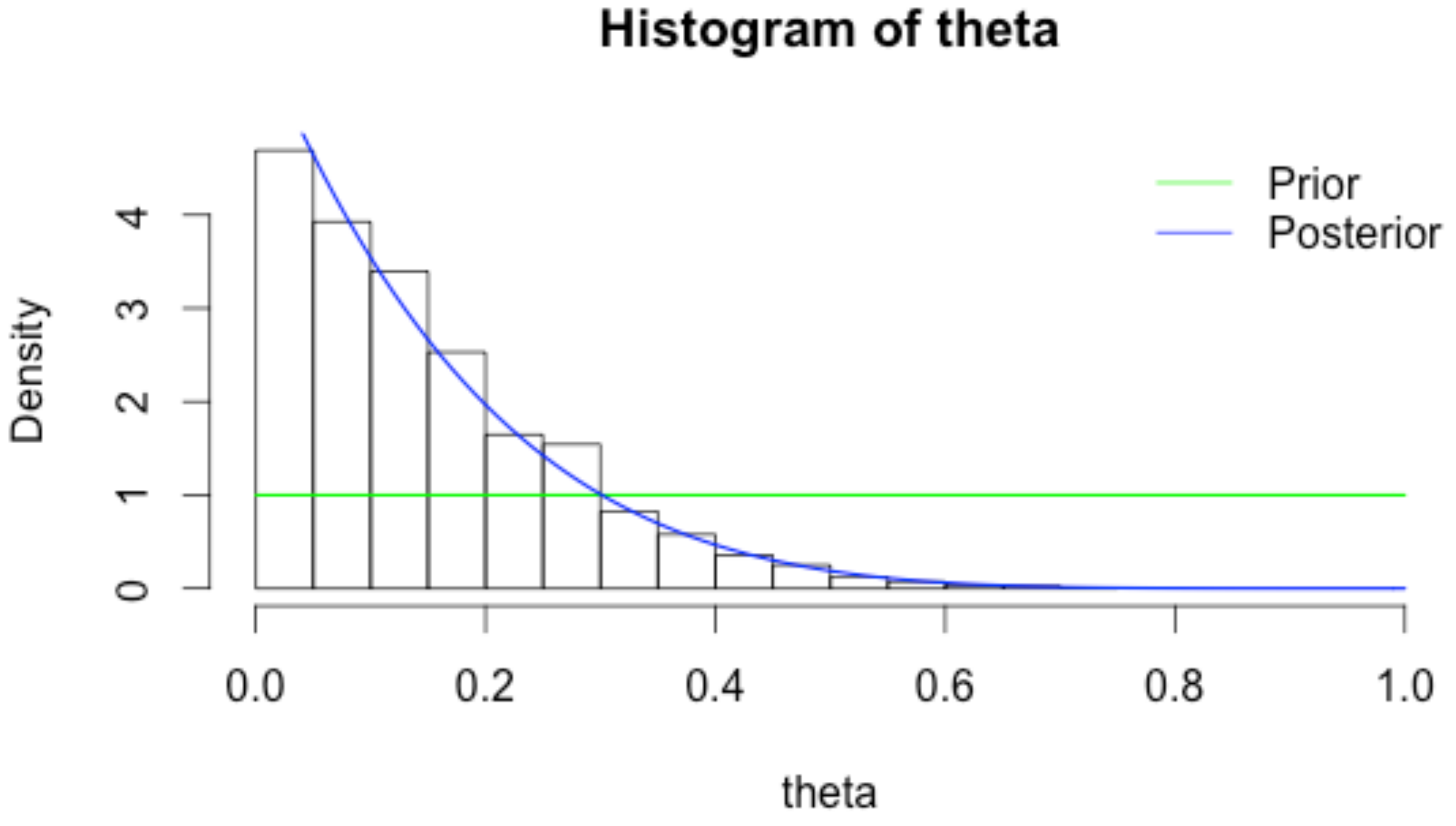
- Posterior predictive mean: $p' = 1/7$=14.3%

# Or use R/Stan

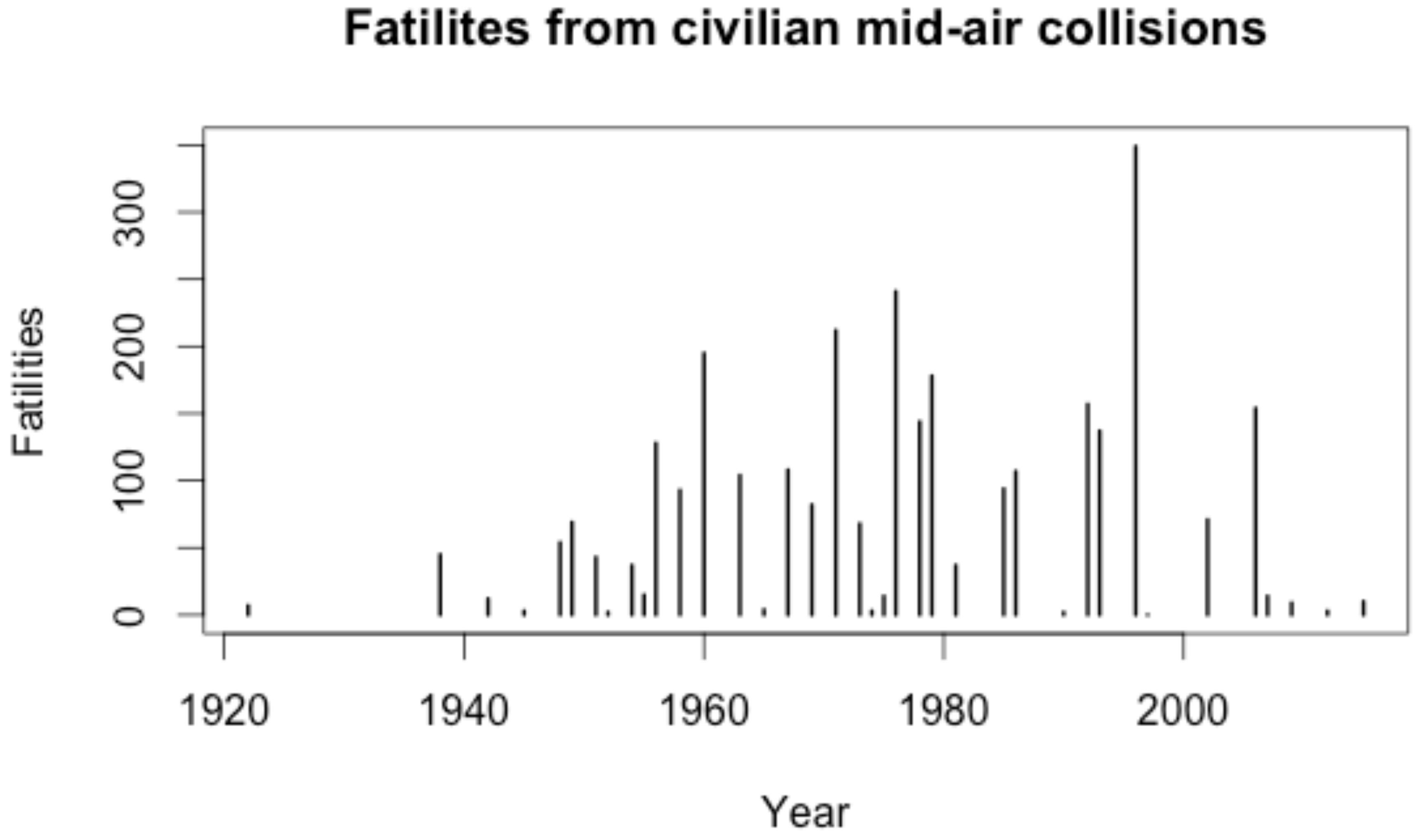
```
library(rstan)
stanmodelcode <- "
data {
  int<lower=0> N;
  int<lower=0, upper=1> y[N];
  }
  parameters {
  real<lower=0, upper=1> theta;
  }
  model {
  theta ~ beta(1, 1);
  for (n in 1:N)
  y[n] ~ bernoulli(theta);
  }
"
fit <- stan(model_code=stanmodelcode, model_name="Longley-Cook",
            data = list(N = 5, y = rep(0,5)))
```

# Review model output

```
fit
## Inference for Stan model: Longley-Cook.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##        mean se_mean   sd 2.5%   25%   50%   75% 97.5% n_eff Rhat
## theta  0.14    0.00 0.12  0.0  0.05  0.11  0.21  0.45  1189    1
## lp__  -3.44    0.02 0.80 -5.7 -3.65 -3.14 -2.93 -2.87  1086    1
```



# Since 1955 there were 11 incidents with more than 100 fatalities



Source: http://en.wikipedia.org/wiki/Mid-air_collision

# Conclusions

- More data is often better
- More thinking time is even better
- Bayesian concepts can turbo charge ‘little’ data/beliefs by borrowing insight from other ‘bigger’ data

# References

Klugman, S. A., Panjer, H. H. & Willmot, G. E. (2004), Loss Models: From Data to Decisions, Wiley Series in Probability and Statistics.

Daniel Kahneman. (2011). Thinking, Fast and Slow. New York : Farrar, Straus and Giroux.

Søren Højsgaard (2012). Graphical Independence Networks with the gRain Package for R. Journal of Statistical Software, 46(10), 1-26. URL http://www.jstatsoft.org/v46/i10/

Computational Actuarial Science with R, Edited by Arthur Charpentier, Chapman and Hall/CRC Reference - 656 Pages

Stan Development Team. 2014. RStan: the R interface to Stan, Version 2.5. http://mc-stan.org/rstan.html.

# The End

Contact:

Markus Gesmann
markus.gesmann@gmail.com
www.magesblog.com