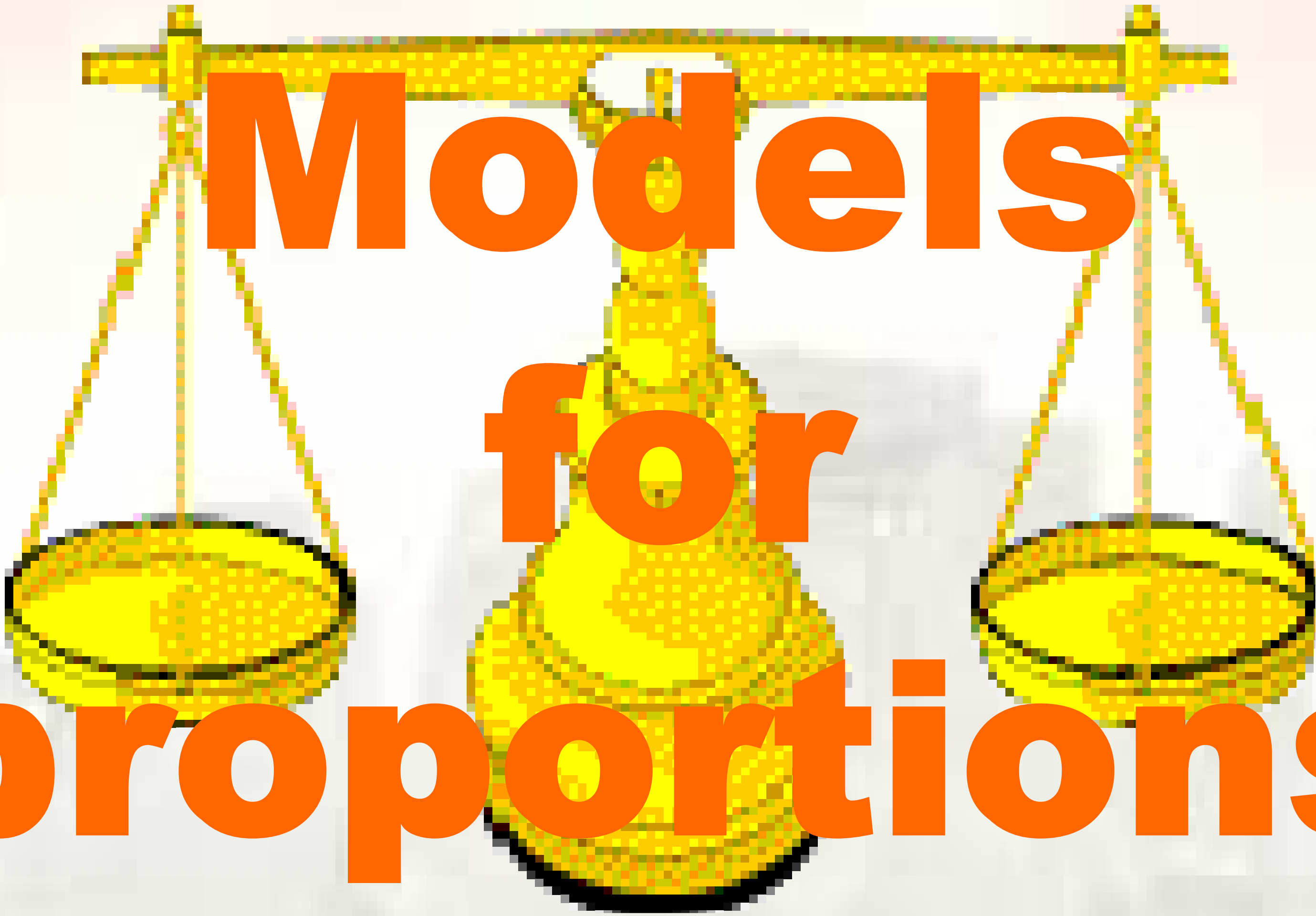
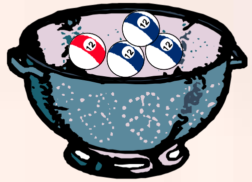
Models for proportions



Model for population

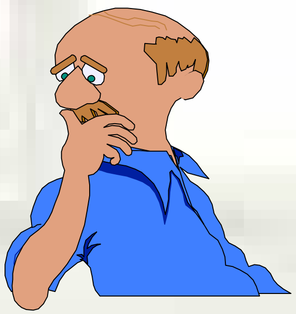


- A model for population can be thought as a bowl with balls labeled according to the possible outcomes of the experiment. The numbers of the various types of balls characterize the model
- a particular model or a subset of models is called a hypothesis
- probability of a hypothesis H is the sum of probabilities of the individual models in H
- a null hypothesis is a model of particular interest-usually one with NO (difference, treatment effect etc.)

"Testing a null hypothesis means finding its posterior probability"

Steps in Bayesian inference

- Specify a set of models
- Assign a prior probability to each model
- Collect data
- Calculate the likelihood P(data|model) of each model
- Use Bayes' rule to calculate the posterior probabilities P(model | data)
- Draw inferences (e.g., predict the next observation)

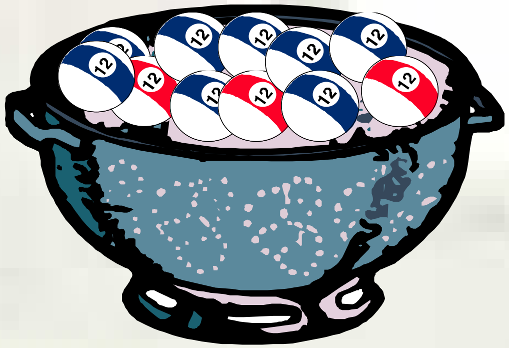


Example

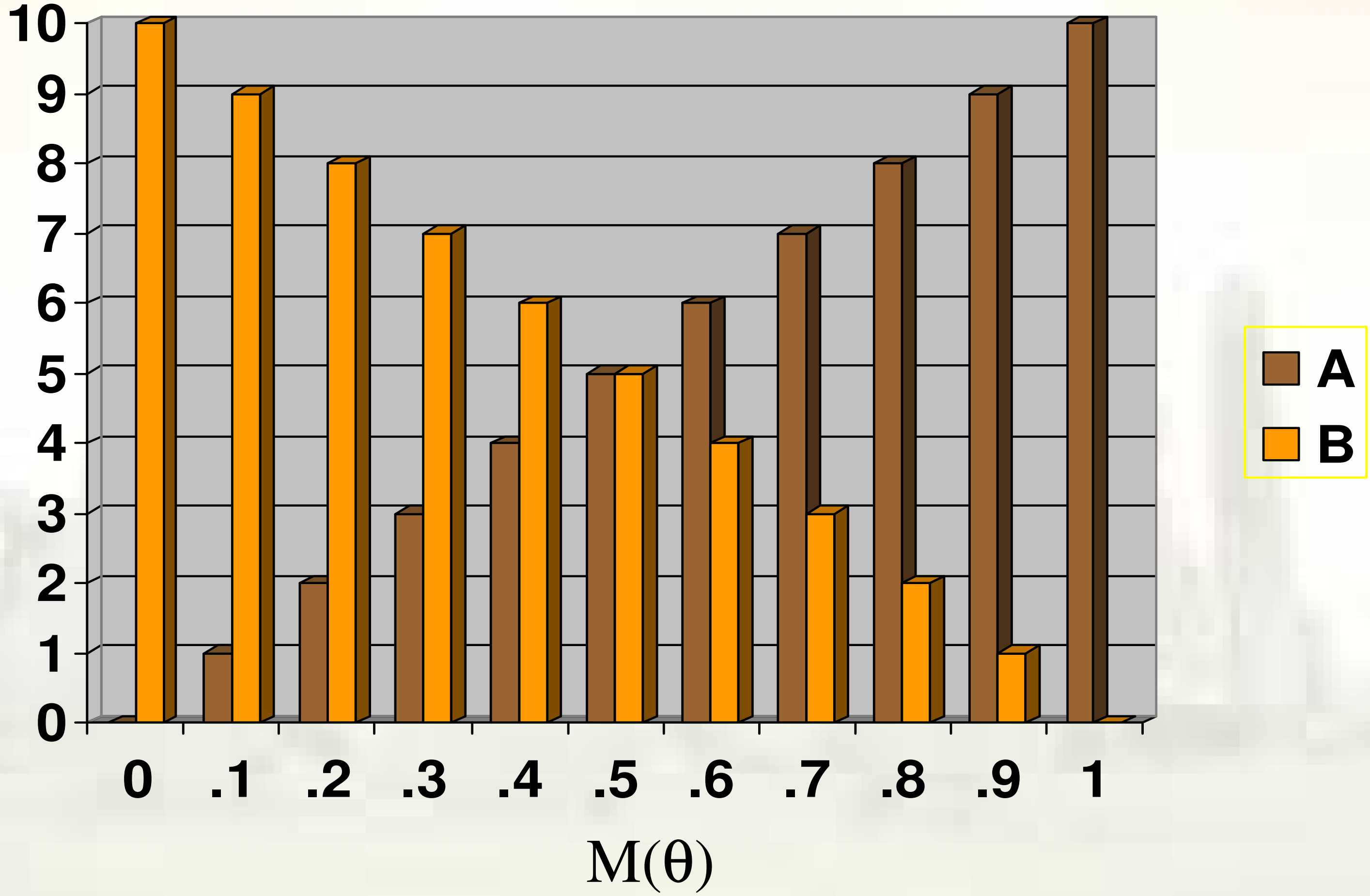
- You are installing WLAN-cards for different machines. You get the WLAN-cards from the same manufacturer, and some of them are faulty.
- We are asking the question: “Is the next WLAN-card we are installing going to work?”
- We are allowed to have background knowledge of these cards (they have been reliable/unreliable in the past, the manufacturing quality has gone up/down etc.)

Assessing models

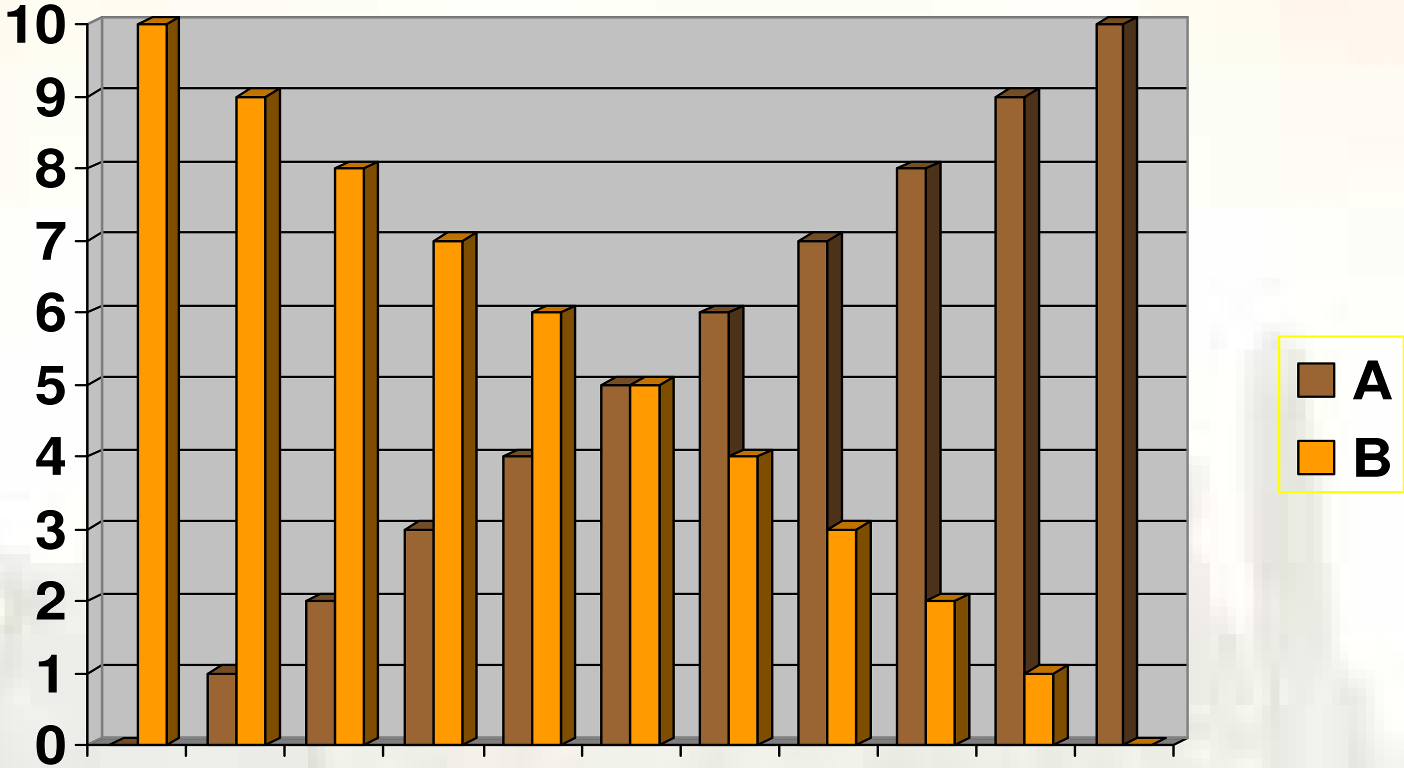
- Let A = "The next WLAN-card is not faulty", and B=~A
- A proportion model can be understood as a bowl with labeled balls (A,B)
- each model M(θ) is characterized by the number of A balls, θ is the proportion (Obs! θ is discrete, i.e., $\theta \in \{0,0.1,0.2,\ldots,1\}$)



Population models

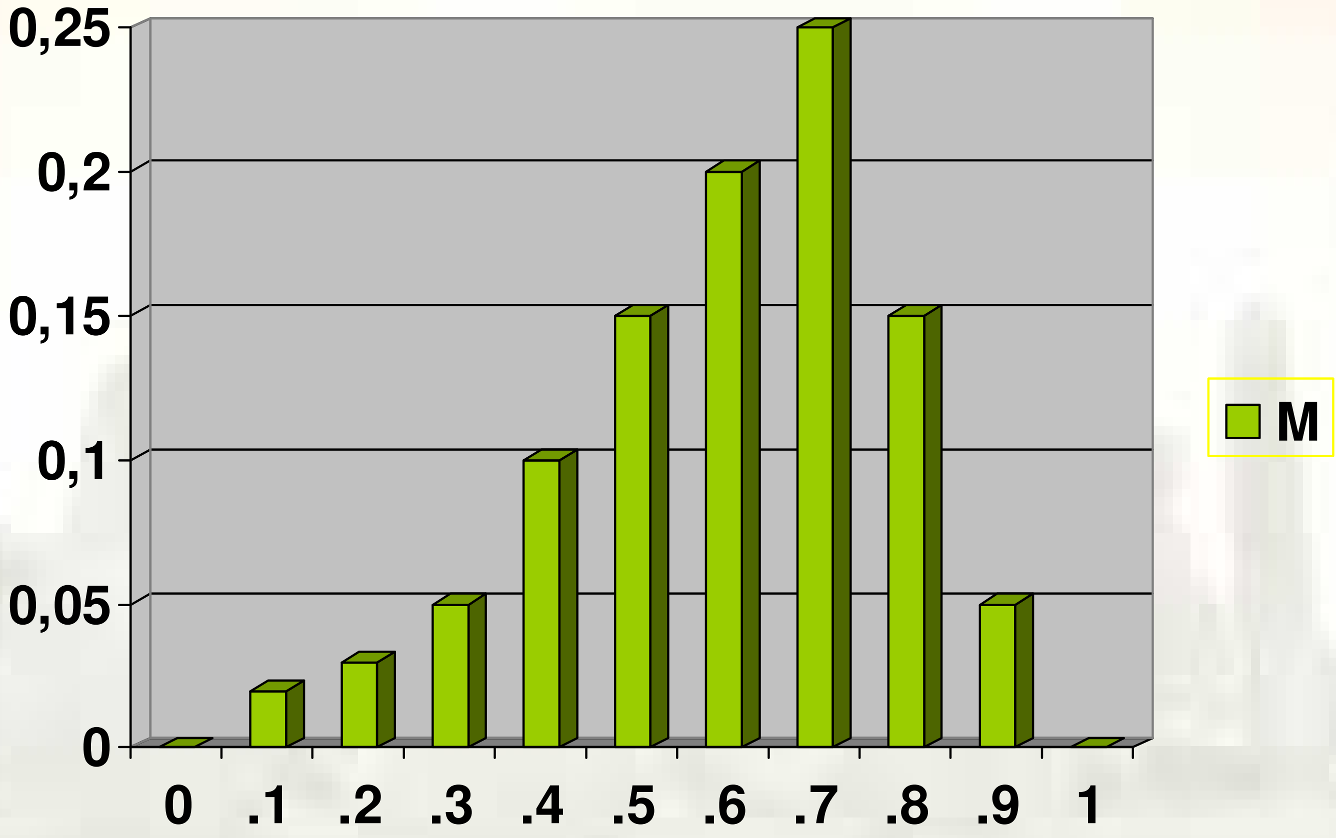


Priors and models



M(θ)	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
P(M(θ))	0	0.02	0.03	0.05	0.1	0.15	0.2	0.25	0.15	0.05	0

Prior distribution



Predictive probability

- What is the probability that the next WLAN-card is not faulty?

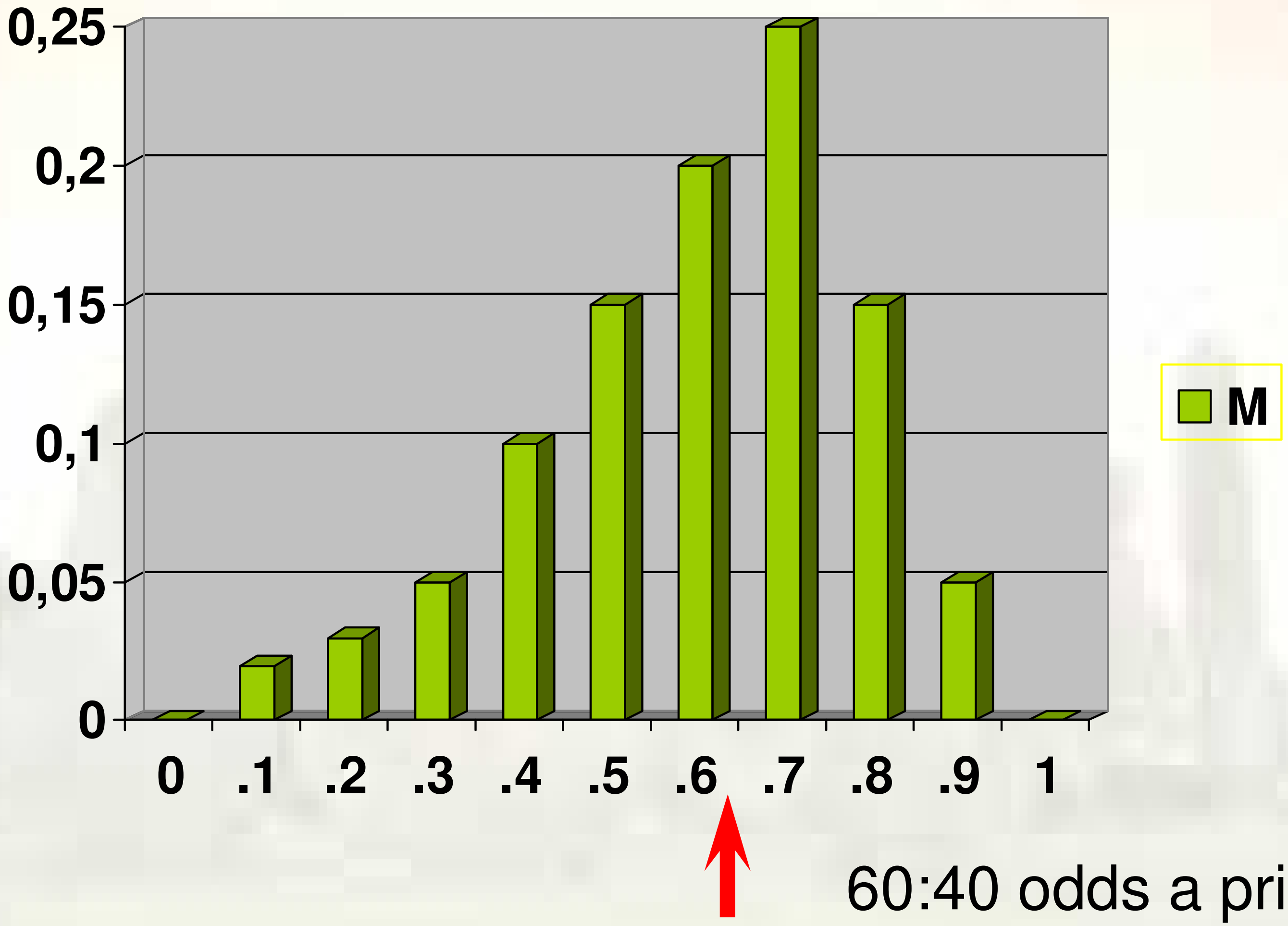
$$
\begin{aligned}
P(A) &= P(A \mid M(0))P(M(0)) + P(A \mid M(0.1))P(M(0.1)) \\
&\quad + \ldots + \mathrm{P}(\mathrm{A} \mid \mathrm{M}(1))\mathrm{P}(\mathrm{M}(1)) \\
&= 0 + 0.002 + 0.006 + 0.015 + \ldots 0 = 0.598
\end{aligned}
$$

Principle of Model averaging

- The previous prediction method is called model averaging, i.e., the uncertainty about the model is taken into account by weighting the predictions of the different alternative models M(θ)

$$P(d \mid M) = \sum_i P(d \mid M(\theta_i), M)\, P(M(\theta_i) \mid M)$$

“Mean or average” model



60:40 odds a priori

Enter more data ...

- Assume that I have installed three WLAN-cards: first was non-faulty, the two latter ones faulty
- what are the updated (posterior) probabilities for the models M(θ)?
- Enter Bayes, for example

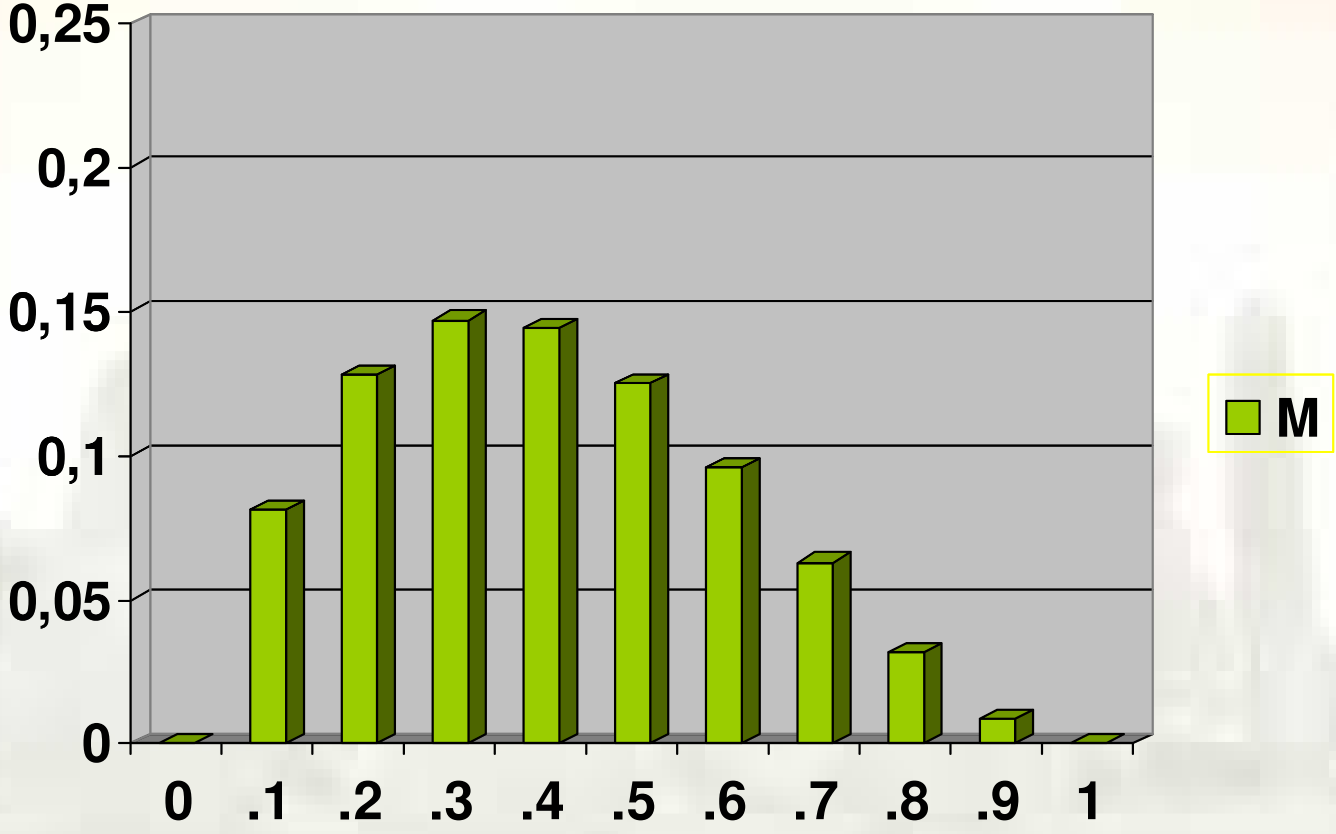
$$P(M(0.6) \mid D) = \frac{P(D \mid M(0.6))\overbrace{P(M(0.6))}^{0.2}}{P(D)}$$

Calculating model likelihoods

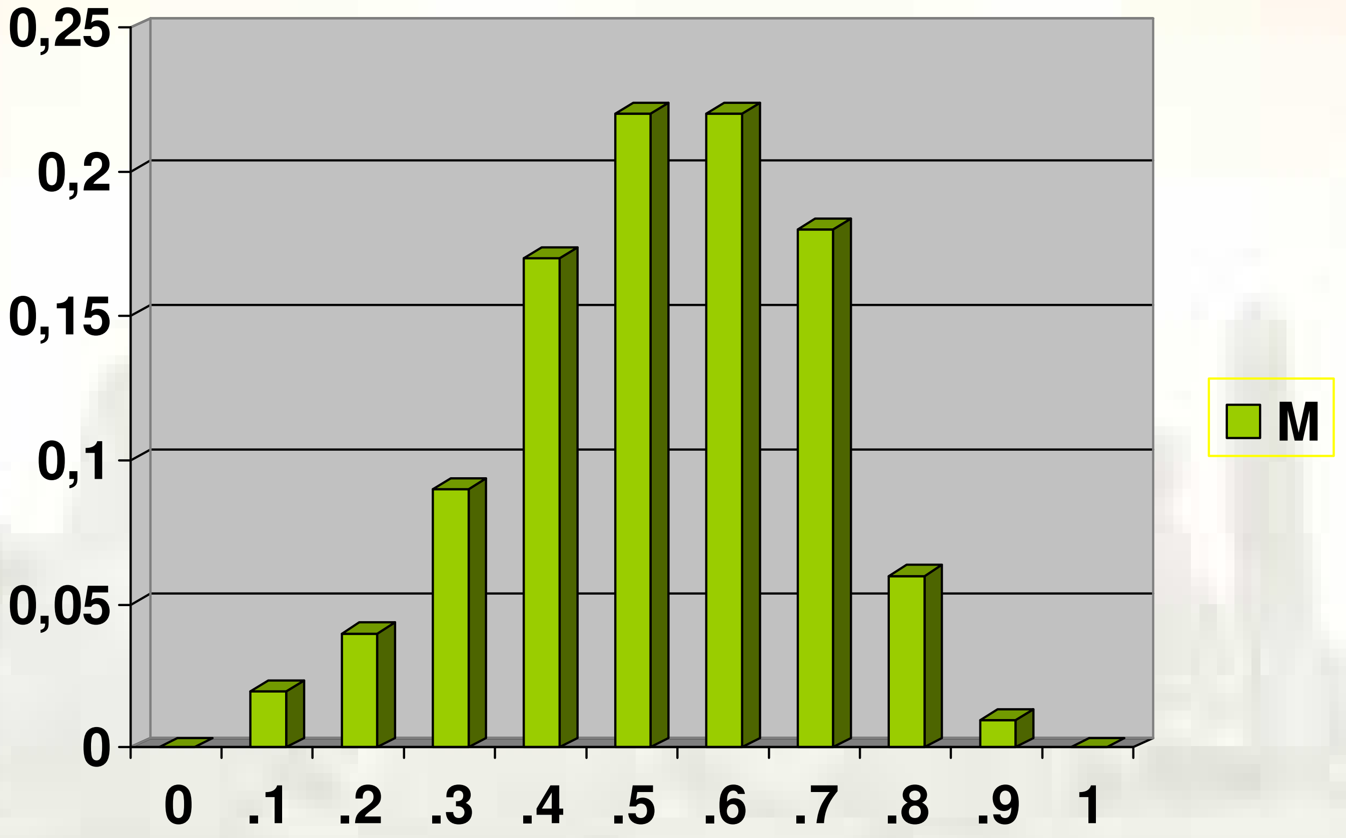
- We assume that the observations are independent given any particular model $M(\theta)$
- P(ABB | M(0.6)) = 0.6 * 0.4 * 0.4 = 0.096
- This is repeated for each model $M(\theta)$

To calculate the *likelihood* of a model, multiply the probabilities of the individual observations given the model

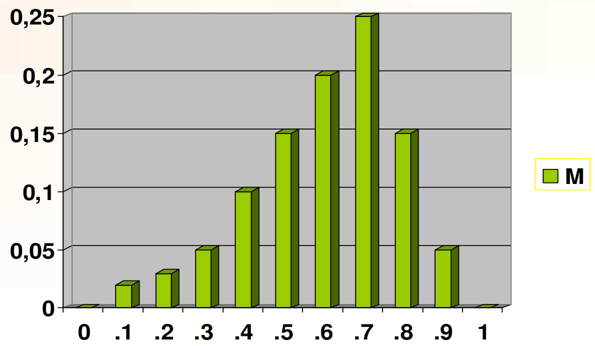
Likelihood histogram P(D|M(θ))



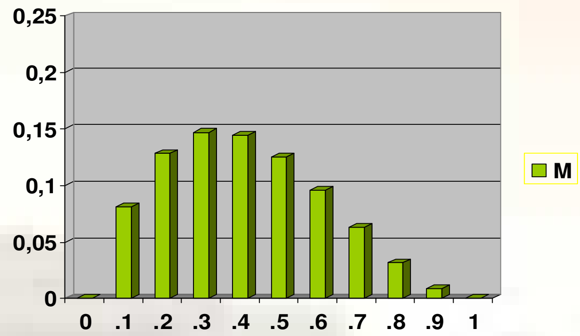
Posterior distribution P(M(θ)|D)



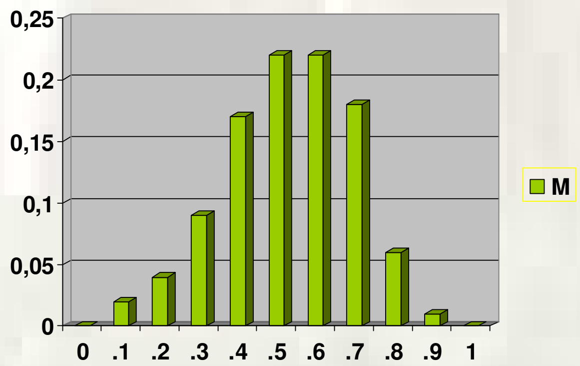
Posterior = likelihood * prior



×



=



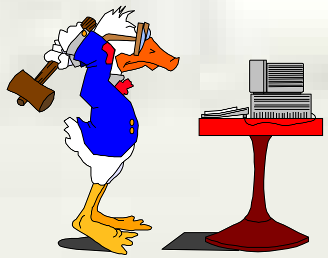
Predictive probability with data D

- with data D the prediction is based on averaging over the models M(θ) weighting by the posterior (instead of the prior used earlier) probability of the models

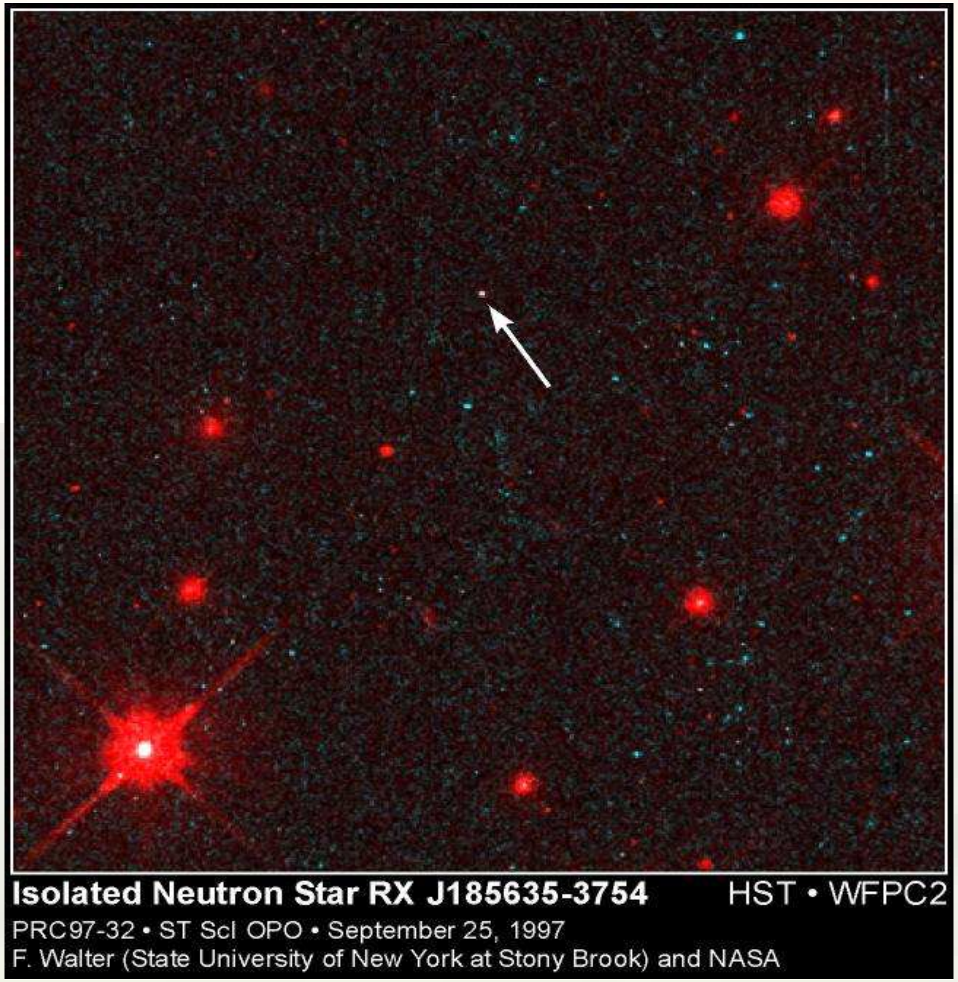
$$P(A \mid D) = \sum_{i \in \{0,0.1,0,\ldots,1\}} P(M(i) \mid D) P(A \mid M(i))$$

How did the probabilities change?

- the posterior distribution is changed: the probability that in general there are more functioning WLAN-cards than malfunctioning cards is down from the prior 65% to 47%
- the predictive probability P(A|D) that the next (fourth) WLAN-card is OK came down from the 60% to 45060/86160 = 52% (the change is not great because the data set is small)



Densities for proportions



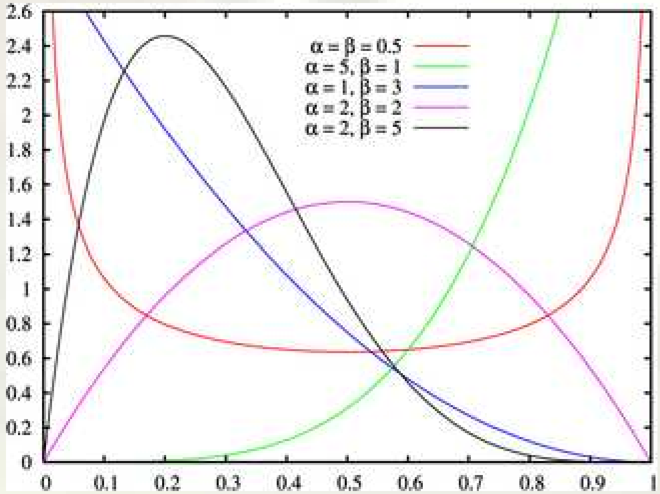
Many models

- a richer set of models allows more precise proportion estimates, but comes with a cost: the amount of calculations necessary increase proportionally
- we can move to consider infinite number of models
  - each model θ is now a point on the interval from [0,1]
  - we get a "smoothed" bar chart called a density $P(\theta)$
  - $\int P(\theta) d\theta = 1$
  - only collections of models can have a probability > 0

Beta Densities

- using densities means that we no longer add probabilities, but calculate areas
- to represent “infinite bar charts” we use curves that approximate the heights of bars
- suppose θ is the success proportion and values $a,b \geq 0$. Density $P(\theta)$ = Beta(a,b) if:

$$P(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$



Updating rule for beta densities

- prior is of form

$$\theta^{a-1}(1-\theta)^{b-1}$$

- assume that you observe s successes and f failures
- in calculating the likelihood whenever s multiply by θ; whenever f multiply by (1- θ). Thus the likelihood is of form

$$\theta^{s}(1-\theta)^{f}$$

- posterior = prior $\times$ likelihood

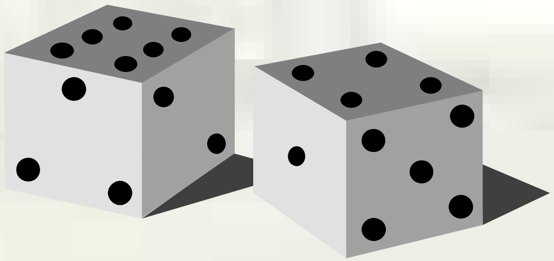
$$\theta^{a-1}(1-\theta)^{b-1}\theta^{s}(1-\theta)^{f} = \theta^{a+s-1}(1-\theta)^{b+f-1}$$

Updating rule for beta densities

- a failure changes the density shape parameter b; a success parameter a

Updating rule for Beta Densities

When the prior is Beta(a,b), and the sufficient statistics of the observed data is s,f, the posterior density is Beta(a+s,b+f)



Predictive probability for beta densities

- Predictive probability of success (A) is

$$P(A \mid a,b) = \int P(A \mid \theta, a,b)\, P(\theta \mid a,b)\, d\theta$$

$$= \int \theta\, P(\theta \mid a,b)\, d\theta = \mathcal{E}(\theta \mid a,b) = a/(a+b).$$

- Hence, one can use a single model θ^* which is the mean of the Beta(a,b) density: $\theta^* = a/(a+b)$
- E.g.: flip a coin 10 times, observe 7 heads ("success"). Assuming a uniform prior Beta(1,1), the posterior for the θ becomes Beta(8,4), and hence the predictive probability of heads is 8/12=2/3.
- Also known as *Laplace's rule of succession*.

Finding beta priors

- assess the probability of success on the first observation (e.g., r(1) = 0.7)
- assume that the first observation was success. Given this information assess the probability of the second success (e.g., r(2) = 0.75)
- So which beta density we choose, i.e., which a and b?

Finding beta priors

$$r(1) = \frac{a}{a+b} \quad \text{and}$$

$$r(2) = \frac{a+1}{a+b+1} \quad \text{gives us}$$

$$a = \frac{r(1)(1-r(2))}{r(2)-r(1)} \quad \text{and} \quad b = \frac{(1-r(1))(1-r(2))}{r(2)-r(1)}$$

e.g., $a = 3.5, b = 1.5$

“Equivalent sample size”

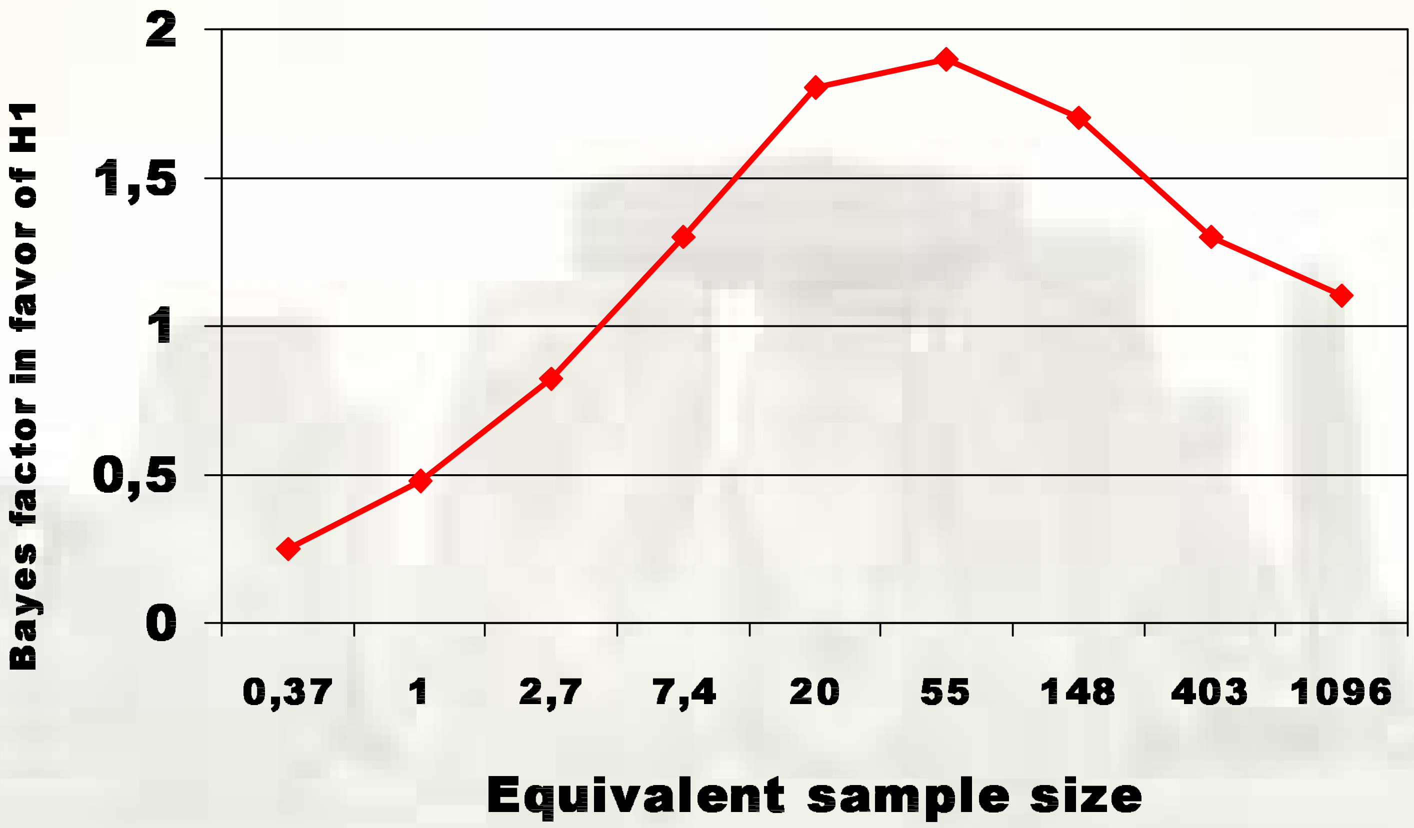
- predictive probabilities change less radically when $a+b$ is large
- interpretation: before formulating prior one has experience of previous observations - thus with $a+b$ one can indicate confidence measured in observations
- called “prior sample size” or “equivalent sample size”
- Beta(1,1) is the uniform prior
- Beta(0.5,0.5) is the Jeffreys prior

Another example

- Toss a coin 250 times, observe D: 140 heads and 110 tails.
- Hypothesis H_0: the coin is fair ($P(\theta=0.5) = 1$)
- Hypothesis H_1: the coin is biased
- Statistics:
  - The P-value is 7%
  - "suspicious", but not enough for rejecting the null hypothesis (Dr. Barry Blight, The Guardian, January 4, 2002)
- Bayes:
  - Let's assume a prior, e.g. Beta(a,a)
  - Compute the Bayes factor

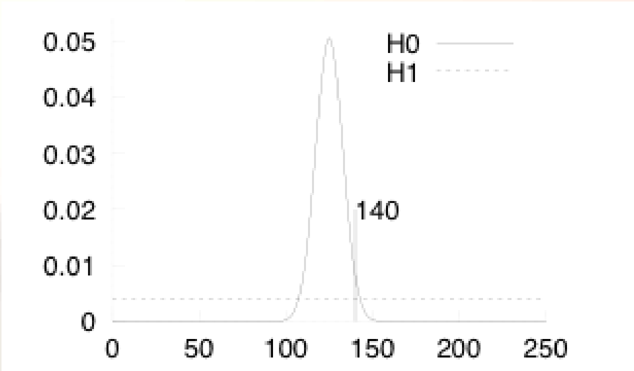
$$\frac{P(D \mid H_1, a)}{P(D \mid H_0)} = \frac{\int P(D \mid \theta, H_1, a) P(\theta \mid H_1, a) d\theta}{\frac{1}{2^{250}}}$$

Equivalent sample size and the Bayes Factor



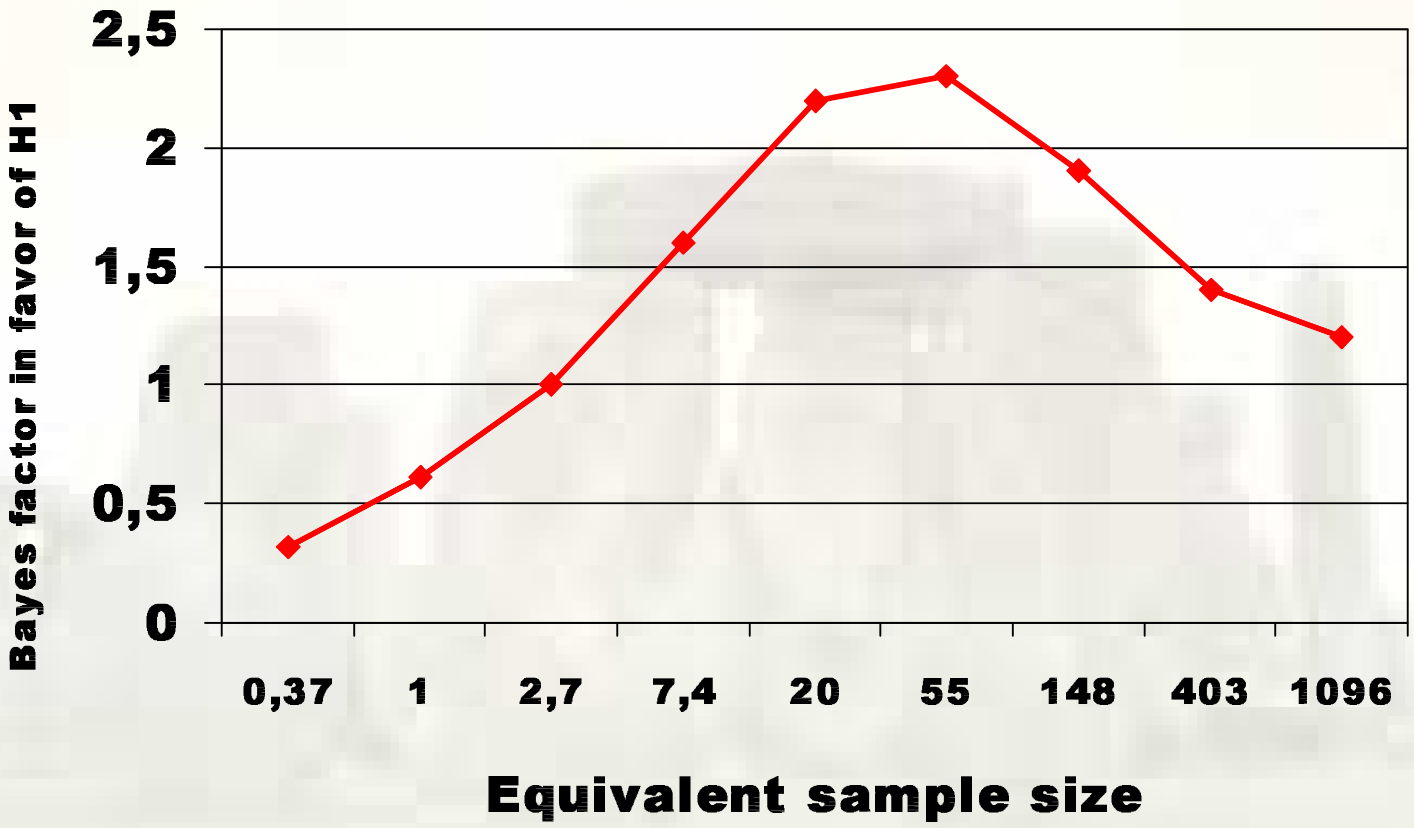
A slightly modified example

- Toss a coin 250 times, observe D = 141 heads and 109 tails.
- Hypothesis H_0: the coin is fair ($P(\theta=0.5) = 1$)
- Hypothesis H_1: the coin is biased
- Statistics:
  - The P-value is 4,97%
  - *Reject the null hypothesis at a significance level of 5%*
- Bayes:
  - Let's assume a prior, e.g. Beta(a,a)
  - Compute the Bayes factor



$$\frac{P(D \mid H_1)}{P(D \mid H_0)} = \frac{\int P(D \mid \theta, H_1, a) P(\theta \mid H_1, a) d\theta}{\frac{1}{2^{250}}}$$

Equivalent sample size and the Bayes Factor (modified example)



Lessons learned

- Classical statistics and the Bayesian approach may give contradictory results
  - Using a fixed P-value threshold is absurd as any null hypothesis can be rejected with sufficient amount of data
  - The Bayesian approach compares models and does not aim at an “absolute” estimate of the goodness of the models
- Bayesian model selection depends heavily on the priors selected
  - However, the process is completely transparent and suspicious results can be criticized based on the selected priors
  - Moreover, the impact of the prior can be easily controlled with respect to the amount of available data
- The issue of determining non-informative priors is controversial
  - Reference priors
  - Normalized maximum likelihood & MDL (see www.mdl-research.org)

On Bayes factor and Occam's razor

- The marginal likelihood (the "evidence") P(D | H) yields a probability distribution (or density) over all the possible data sets D.
- Complex models can predict well many different data sets, so they need to spread the probability mass over a wide region of models