Bayesian probability: **P**

State of the World: **X**

P(**X** | your information **I**)

First example: bag of balls

- Every probability is conditional to your background knowledge ”I”: P(A | I)
- What is the (your) probability that there are r red balls in a bag? (Assuming N balls which can be red/white)
- Before any data, you might select your **prior probability** as $P(r)=1/(N+1)$ for all possible r. (0,1,2,…,N).
  - Here r is the unknown parameter, and your data will be the observed balls that will be drawn.

First example: bag of balls

- Given that there are i/N red balls, you might say: the probability of picking 'blindly' one red ball is
  $P(X=red \mid i/N) = i/N$
- This is your (subjective) model choice.
- Calculate **posterior probability**:
  $P(r=i/N \mid X=red)$

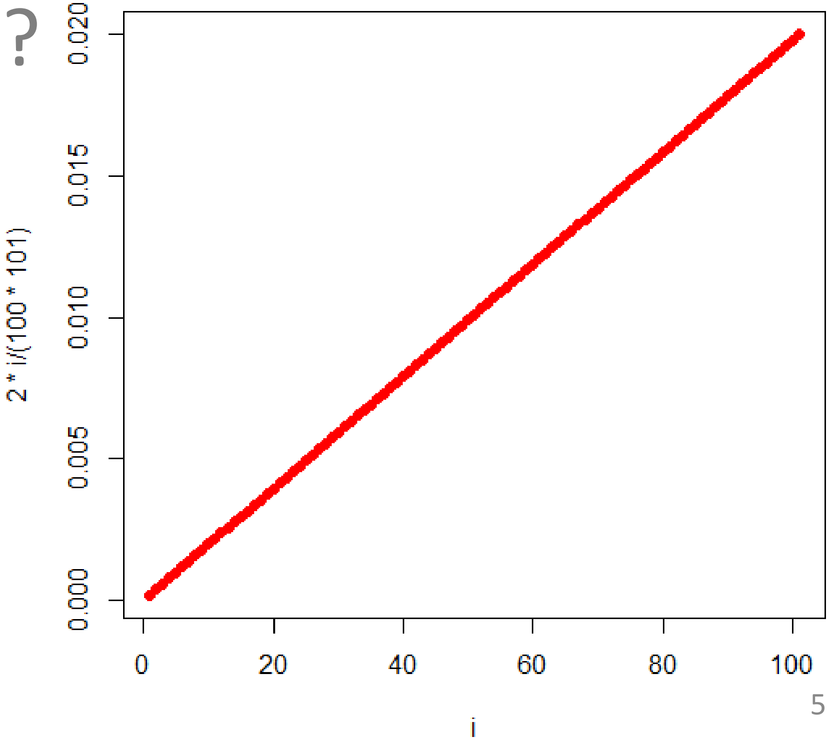
First example: bag of balls

- Remember some probability calculus:
- P(A,B)=P(A|B)P(B)=P(B|A)P(A)=P(B,A)
- Joint probability in this example:
- P(X=red,r=i/N) = (i/N)*(1/(N+1))
- Calculate: P(r=i/N | X=red)
  = (i/N)*(1/(N+1)) / P(X=red)
- P(X=red) is just normalizing constant, i.e.

$$P(X = red) = \sum_{i=0}^{N} P(X = red \mid r = i/N) P(r = i/N) = 1/2$$

First example: bag of balls

- The posterior probability is therefore:
- P(r=i/N | X=red) = 2i/(N*(N+1))
- What have we learned from the observation ”X=red”?
- Compare with the prior probability.



First example: bag of balls

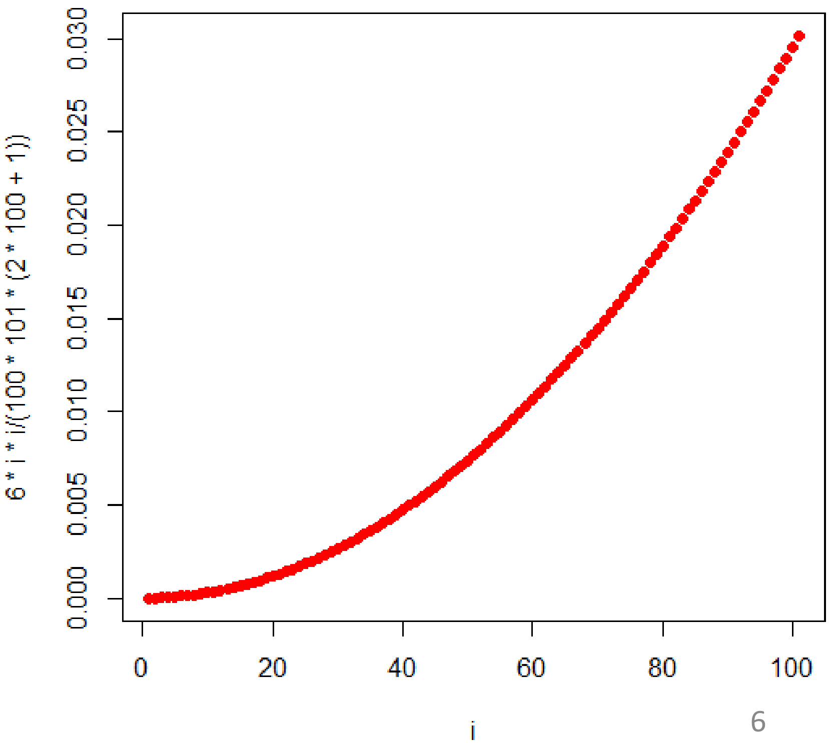
- Our new prior is: 2i/(N*(N+1))
- After observing two red balls ”X=2red”:
- Now: P(r=i/N | X=2)

  $= (i/N) * 2i/(N(N+1))/c$

  $= 2i^2/(N^2(N+1))/c$
- Normalizing constant

  $c = (2N+1)/3N$
- So: P(r=i/N | X=2)

  $= 6i^2/(N(N+1)(2N+1))$



First example: bag of balls

- The result is the same if
  - Start with original prior, + use the probability of observing two red balls
  - Start with the posterior we got after observing one red ball, + use the probability of observing one red ball (again)
- The model would be different if we assume that balls are not replaced in the bag.

First example: bag of balls

- The prior (and posterior) probability P(r) can be said to describe epistemic uncertainty.
- The conditional probability P(X|r) can be said to describe aleatoric uncertainty.
- Where do these come from?
  - Background information.
  - Model choice.

Elicitation of a prior from an expert

- P(A) should describe the expert's beliefs.
- Consider two options:
  - You'll get €300 if "A is true"
  - You'll get a lottery ticket knowing n out of 100 wins €300.

  Which option do you choose?

  $n_{small}/100 < P(A \mid your) < n_{large}/100$

  Can find out: $n/100 \approx P(A \mid your)$

Elicitation of a prior from an expert

- Also, in terms of odds $w = P(A)/(1-P(A))$, a fair bet is such that

  $$P(A)wR + (1-P(A))(-R) = 0$$

  Find out $P(A) = 1/(1+w)$

- **Probability densities more difficult to elicit.**
- **Multivariate densities even more difficult.**
- **Psychological biases.**

Elicitation of a prior from an expert

- Assume we have elicited densities $p_i(x)$ from experts i=1,…,N.
- **Combination?**

Mixture density $$p(x) = \sum_{i=1}^{N} p_i(x) \times \frac{1}{N}$$

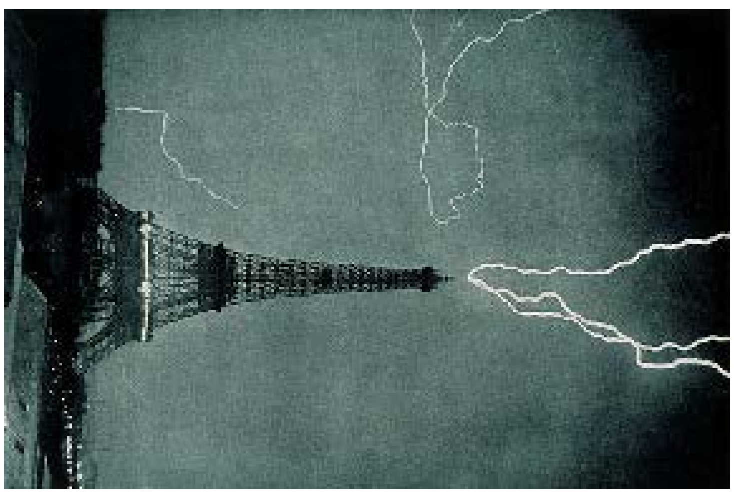
Product of densities: $$p(x) = \prod_{i=1}^{N} p_i(x)^{1/N} / c$$

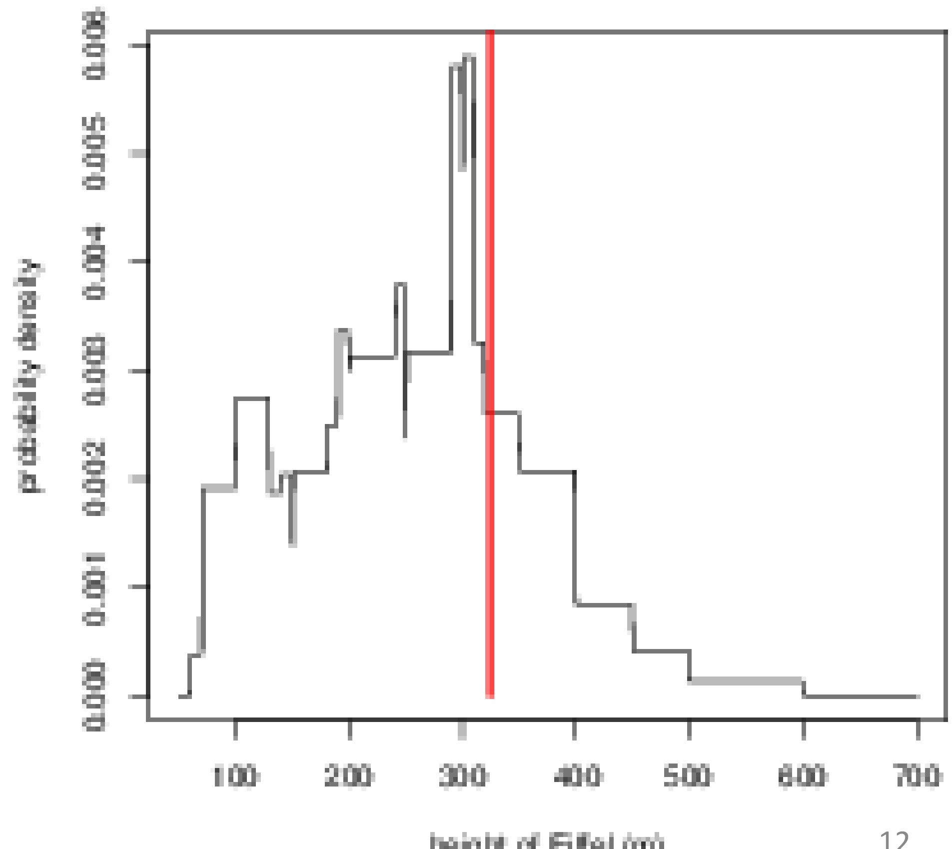
(needs normalizing constant c)

The height of Eiffel?

- What's your minimum and maximum?

→ **$p_i = U(min_i, max_i)$**





Choice of prior

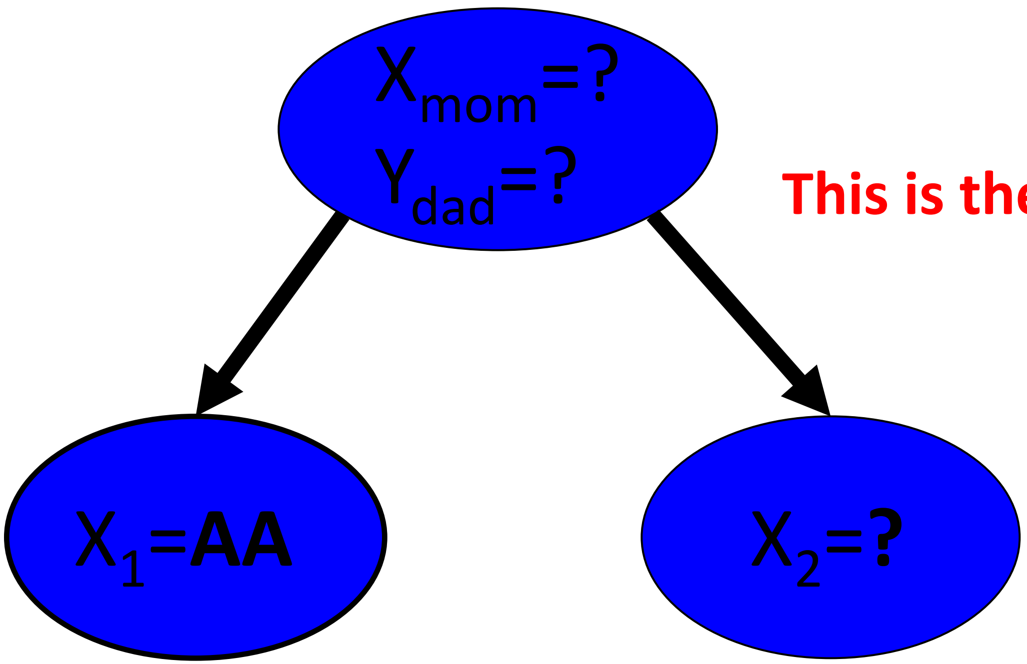
- Subjective expert knowledge can be important
  - When we have little data.
  - When it is the only source of information.
  - When data would be too expensive.
  - Difficult problems never have sufficient data…
- Alternatively: uninformative, ’flat’ priors.
- **’Objective Bayes’ & ’Subjective Bayes’**

An example from school book genetics

- Assume parents with unknown genotypes:
  - **Aa, aa or AA**.

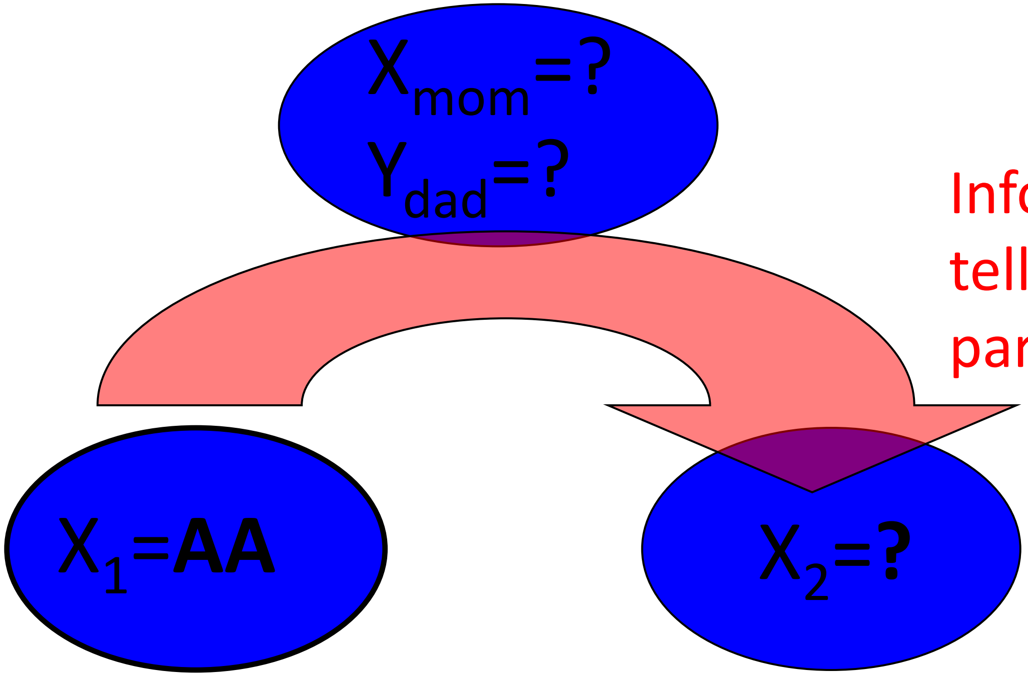
- Assume a child is observed to be of type **AA**.
- **Question1**: now what is the probability for the genotypes of the parents?
- **Question2**: what is the probability that the next child will also be of type **AA**?

- Graphically: there is a conditional probability for the genotype of each child, ***<u>given</u>*** the type of parents:



This is the "data model"

- Now, given the prior AND the observed child, we calculate the probability of the 2nd child:



Information about 1st child tells something about the parents, hence about the 2nd child.

The ***posterior*** probability is **1/4** for each of the parental combinations:

[AA,AA] , **[Aa,Aa]** , **[AA,Aa]** , **[Aa,AA]**

This Results to: **P(AA)=9/16** for the 2nd child.
Compare this with prior probability: **P(AA)=1/4.**
New evidence changed this.

- Using Bayes: $P(A|B) = \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)}$

- the ***posterior*** probability for the parents can be calculated as:

$$P(\mathrm{X}_{\mathrm{mom}}, \mathrm{X}_{\mathrm{dad}} | \mathrm{X}_1 = AA) = \frac{P(\mathrm{X}_1 = AA | \mathrm{X}_{\mathrm{mom}}, \mathrm{X}_{\mathrm{dad}}) P(\mathrm{X}_{\mathrm{mom}}, \mathrm{X}_{\mathrm{dad}})}{\sum_{\mathrm{X}_{\mathrm{mom}}} \sum_{\mathrm{X}_{\mathrm{dad}}} P(\mathrm{X}_1 = AA | \mathrm{X}_{\mathrm{mom}}, \mathrm{X}_{\mathrm{dad}}) P(\mathrm{X}_{\mathrm{mom}}, \mathrm{X}_{\mathrm{dad}})}$$

- This describes our final degree of uncertainty.

- The ***posterior*** probability is **1/4** for each of the parental **combinations**:

[AA,AA] , [Aa,Aa] , [AA,Aa] , [Aa,AA]

- Notice, "**aa**" is no longer a possible type for either parent. The prediction for the next child is thus:

$$P(X_2 = AA \mid X_1 = AA) = \sum_{X_{mom} X_{dad}} P(X_2 = AA \mid X_{mom} X_{dad}) \underbrace{P(X_{mom} X_{dad} \mid X_1 = AA)}_{Posterior}$$

- Resulting to: **9/16**
- Compare this with prior probability: **P(AA)=1/4**

- The previous example had all the elements that are essential.
  - **The same idea is just repeated in various forms.**

Binomial model

- Recall Bayes' original example.
- $X \sim \text{Binomial}(N,\theta)$
- $p(\theta)$ = prior density.
  - $U(0,1)$
  - $\text{Beta}(\alpha,\beta)$
- Find out $p(\theta|X)$

Binomial model

- **Posterior density: P(θ | X)=P(X|θ)P(θ)/c**
  - Assuming uniform prior:

$$p(\theta \mid x) = \binom{N}{x} \theta^{x} (1-\theta)^{N-x} 1_{\{0<\theta<1\}}(\theta) / c$$

  - Take a look at this as a function of θ, with N, x, and c as fixed constants.
  - What probability density function can be seen? Hint: compare to beta-density.

$$p(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Binomial model

- The posterior can be written, up to a constant term as

$$p(\theta \mid N, x) \propto \theta^{x+1-1}(1-\theta)^{N-x+1-1}$$

  - Same as beta(x+1,N-x+1)
  - If the uniform prior is replaced by beta(α,β), we get beta(x+α,N-x+β)

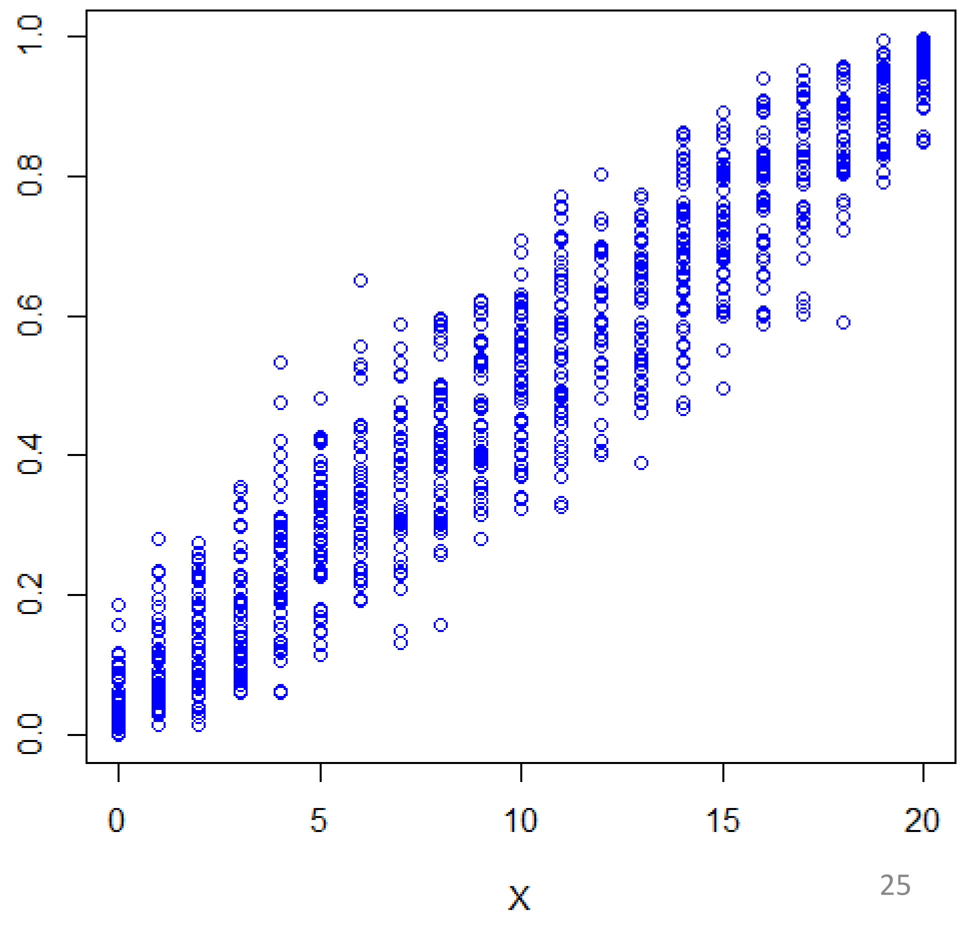
Binomial model

- The uniform prior corresponds to having two ’pseudo observations’: one red ball, one white ball.
- The posterior mean is (1+X)/(2+N)
  - Or: $(\alpha+X)/(\alpha+\beta+N)$
  - Can be expressed as $$w\frac{\alpha}{\alpha+\beta}+(1-w)\frac{X}{N}$$

    With $w = (\alpha+\beta)/(\alpha+\beta+N)$
  - See what happens if $N \rightarrow \infty$, or if $N \rightarrow 0$.

Binomial model

- Simulated sample from the joint distribution $p(\theta,X)= P(X|N,\theta)p(\theta)$
- See $P(X|N,\theta)$ and $p(\theta|X)$ in the Fig.



Binomial model

- The binomial distribution (likelihood function) and the beta-prior are said to be conjugate.
- Conjugate choice of prior leads to closed form solutions. (Posterior density is in the same family as prior density).
- Can also interpret conjugate prior as ’pseudo data’ in comparison with real data.
- Only a few conjugate solutions exist!

Binomial model & priors

- The uniform prior U(0,1) for θ was 'uninformative'. In what sense?
- What if we study the density of θ^2 or $\log(\theta)$, assuming $\theta \sim U(0,1)$?
- Jeffreys' prior is uninformative in the sense that it is transformation invariant:

$$p(\theta) \propto J(\theta)^{1/2}$$ with $$J(\theta) = E[(\frac{d \log(P(X \mid \theta))}{d\theta})^2 \mid \theta]$$

Binomial model & priors

- J(θ) is known as 'Fisher information for θ'
- With Jeffreys' prior for θ we get, for any one-to-one smooth transformation φ=h(θ) that:

Transformation of variables rule

Jeffreys'

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| \propto \sqrt{E[(\frac{d\log(L)}{d\theta})^2(\frac{d\theta}{d\phi})^2]}$$

$$= \sqrt{E[(\frac{d\log(L)}{d\phi})^2]} = \sqrt{J(\phi)}$$ where L = P(X|parameter)

Binomial model & priors

- For the binomial model, Jeffreys' prior is Beta(1/2,1/2).
- But in general:
  - Jeffreys' prior can lead to improper densities (integral is infinite).
  - Difficult to generalize into higher dimensions.
  - Violates likelihood principle which states that inferences should be the same when the likelihood function is the same.

Binomial model & priors

- Also: Haldane's prior Beta(0,0) is uninformative.
  - (How? Think of 'pseudo data'…)
  - But is **improper**.
- *Can a prior be improper density?*
  - Yes, but! - the likelihood needs to be such that the posterior still integrates to one.
  - With Haldane's prior, this works only when the binomial data X is either >0 or <N.

Binomial model & priors

- For the binomial model $P(X|\theta)$, when computing the posterior $p(\theta|X)$, we have at least 3 different uninformative priors:

- $p(\theta)=U(0,1)=Beta(1,1)$ Bayes-Laplace
- $p(\theta)=Beta(1/2,1/2)$ Jeffreys'
- $p(\theta)=Beta(0,0)$ Haldane's

- Each of them is uninformative in different ways!
- Unique definition for **uninformative** does not exist.

Binomial model & priors

- example: estimate the mortality

THIRD DEATH

The expanded warning came as Yosemite announced that a third person had died of the disease and the number of confirmed cases rose to eight, all of them among U.S. visitors to the park.

Ok, it's a small data,

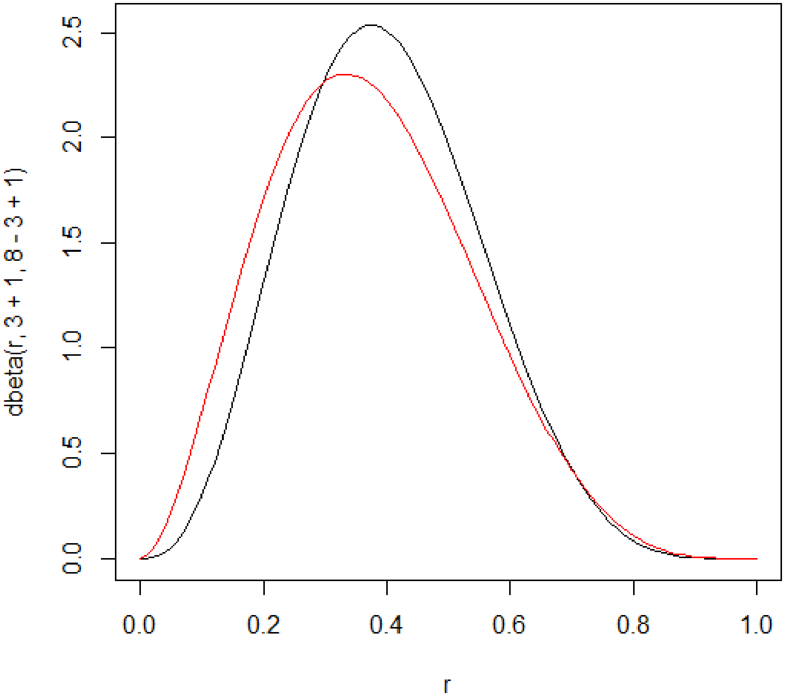
but we try:

with uniform prior:

p(r | data)=beta(3+1,8-3+1).

Try also other priors.

(Haldane's in red →)



Binomial model & N?

- In previous slides, N was fixed (known). We can also think situations where θ is known , X is known, but N is unknown.
- Exercise: solve $P(N \mid \theta, X) = P(X \mid N, \theta)P(N)/c$ with suitable choice of prior.
  - Try e.g. discrete uniform over a range of values.
  - Try e.g. $P(N) \propto 1/N$

- With Bayes rule we can compute probabilities of any unknowns, given the knowns & prior & likelihood (model).

Poisson model

- Widely applicable: counts of disease cases, accidents, faults, births, deaths over a time, or within an area, etc…

$$P(X \mid \lambda) = \frac{\lambda^X}{X!} e^{-\lambda}$$

- λ = Poisson intensity = E(X).
- Aim to get: p(λ|X)
- Bayes: p(λ|X) = P(X|λ)p(λ)/c

Poisson model

- Conjugate prior? Try Gamma-density:

$$p(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

- Then:

$$p(\lambda \mid X) = \frac{\lambda^{X}}{X!} e^{-\lambda} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} / c$$

- Simplify expression, what density you see? (up to a normalizing constant).

Poisson model

- Posterior density is Gamma(X+α,1+β).
- Posterior mean is (X+α)/(1+β)
- Can be written as weighted sum of 'data mean' X and 'prior mean' α/β.

$$\frac{1}{1+\beta}X + \frac{\beta}{1+\beta}\frac{\alpha}{\beta}$$

Poisson model

- With a set of observations: $X_1,\ldots,X_N$:

$$P(X_1,\ldots,X_N \mid \lambda) = \prod_{i=1}^{N} \frac{\lambda^{X_i}}{X_i!} e^{-\lambda}$$

- And with the Gamma(α,β)-prior we get: Gamma($X_1+\ldots+X_N+\alpha, N+\beta$).

- Posterior mean $\frac{1}{N+\beta}\sum_{i=1}^{N} X_i + \frac{\beta}{N+\beta}\frac{\alpha}{\beta}$

- What happens if $N \rightarrow \infty$, or $N \rightarrow 0$?

Poisson model

- Gamma-prior can be informative or uninformative. In the limit $(\alpha,\beta) \rightarrow (0,0)$, posterior $\rightarrow$ Gamma($X_1+\ldots+X_N$,N).

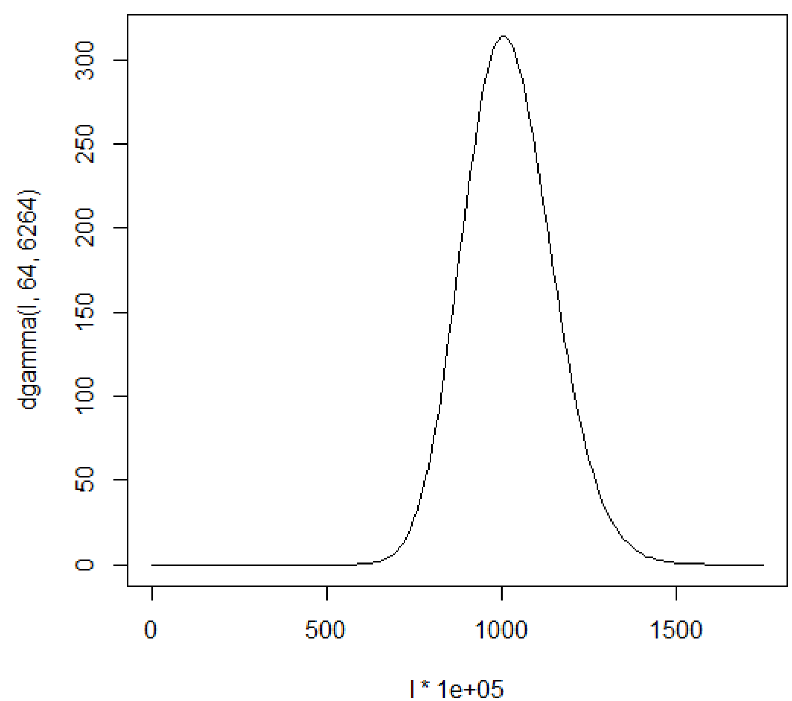
- Compare the conjugate analysis with Binomial model. Note similarities.

Poisson model

- Parameterizing with exposure
  - Type of problems: rate of cases per year, or per 100,000 persons per year.
  - Model: $X_i \sim \text{Poisson}(\lambda E_i)$
  - E_i is **exposure**, e.g. population of the i^{th} city (in a year).
  - X_i is observed number of cases.

Poisson model

- Example: 64 lung cancer cases in 1968-1971 in Fredericia, Denmark, population 6264. Estimate incidence per 100,000?
- $P(\lambda | X,E)$
  $= \text{gamma}(\alpha+\Sigma X_i, \beta+\Sigma E_i)$
- With non-informative prior, X=64,E=6264, we get gamma(64,6264), (plot: $10^5\ \lambda$)



Exponential model

- Applicable for event times, concentrations, positive measurements,…

$$p(X \mid \theta) = \theta e^{-\theta X}$$

- Mean $E(X) = 1/\theta$
- Aim to get $P(\theta|X)$, or $P(\theta|X_1+\ldots+X_N)$.
- Conjugate prior $\text{Gamma}(\alpha,\beta)$
- Posterior: $\text{Gamma}(\alpha+1,\beta+X)$ or $\text{Gamma}(\alpha+N,\beta+X_1+\ldots+X_N)$.

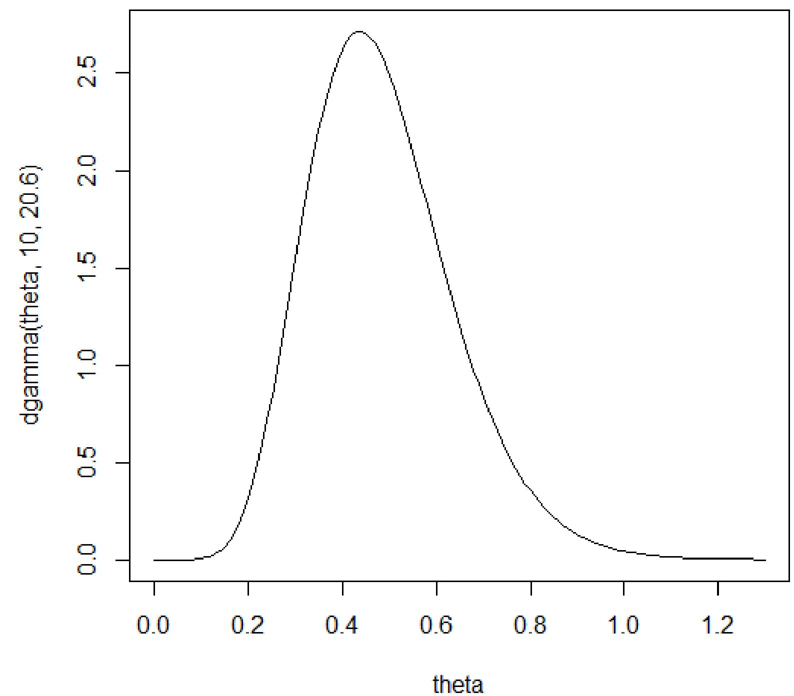
Exponential model

- Posterior mean is $(\alpha+N)/(\beta+X_1+\ldots+X_N)$
- What happens if $N \rightarrow \infty$, or $N \rightarrow 0$?
- Uninformative prior $(\alpha,\beta) \rightarrow (0,0)$

- Similarities again.

Exponential model

- Example: life times of 10 light bulbs were T = 4.1, 0.8, 2.0, 1.5, 5.0, 0.7, 0.1, 4.2, 0.4, 1.8 years. Estimate the failure rate? (true=0.5)
- $T_i \sim \exp(\theta)$
- Non-informative prior gives $p(\theta|T)$ = gamma(10,20.6).
- Could also parameterize with $1/\theta$ and use inverse-gamma prior.



Exponential model

- Some observations may be censored, so we know only that $T_i < c_i$, or $T_i > c_i$
- The probability for the whole data is then of the form:
- $P(\text{data} \mid \theta) =$

$$\prod P(T_i \mid \theta) \prod P(T_i < c_i \mid \theta) \prod P(T_i > c_i \mid \theta)$$

- Here we need cumulative probability functions, but the Bayes theorem still applies, just more complicated.

Binomial, Poisson, Exponential

- The simplest one-parameter models.
- Conjugate priors available.
- Prior can be seen as ’pseudo data’ comparable with actual data.
- Easy to see how the new data update the prior density to posterior density.
- Posterior means, variances, modes, quantiles can be used to summarize.