Background (chapter 1)

Slide 1

- Terms and notation
- Bayes' rule, sum and product rules
- Predictive distribution
- Probability as a measure of uncertainty
- Subjectivity vs. objectivity
- Direct simulation
- Inverse-cdf method

Single-parameter models (chapter 2)

Slide 2

- Terms and notation
- Binomial
  - what kind of data, equation, parameters, conjugate prior
- Normal
  - what kind of data, equation, parameters, conjugate priors
  - (no need to derive posterior distributions in exam)
- Summarizing posterior distributions
  - mean, median, deviation, variance, quantiles, intervals, HPD
- conjugate prior vs. non-conjugate prior
  - pros and cons, eg. effect to computation
- Informative prior vs. non-informative prior
  - pros and cons

Single-parameter models

- Binomi — Jacob Bernoulli (1654-1705), Bayes (1702-1761)
- Normaali — de Moivre (1667-1754), Laplace (1749-1827), Gauss (1777-1855)
- Poisson — Poisson (1781-1840)
- Exponential, Double exponential — Laplace (1749-1827)

Slide 3

Binomial

- Data $y_1, \ldots, y_n$, where y_i is either 0 or 1
- Assume exchangeability of the events
  - order of the events does not matter
  - thus all we need to know is number of 0's and 1's
    - Bernoulli's urn
    - e.g. colored chips in the bag
    - e.g. how many girls and boys were born

Slide 4

- Exchangeability discussed in more detail in ch 5

Binomial

- Assume same number of colors, probability of white 0.5
- Even proportion, pr. of n whites $0.5, \ldots, 0.5 = 0.5^n$
- Uneven proportion, n chips, y white, in some order $\theta * (1-\theta) \ldots = \theta^y (1-\theta)^{n-y}$

Slide 5

- Any order, ie. sum over different permutations $p(y|\theta, n, M) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$
  - data can be summarized by y and n

Binomial

- Assuming binomial model and parameter θ, we can proceed *as if events were independent and identically distributed given model* M *and parameter* θ

$$p(y|\theta, n, M) = \text{Bin}(y|n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Slide 6

Binomial: posterior of θ

- Bayes' rule

$$p(\theta|y,n,M) = \frac{p(y|\theta,n,M)p(\theta|n,M)}{p(y|n,M)}$$

- Let's start with a simple prior

$$p(\theta|n,M) = p(\theta|M) = 1, \text{ if } 0 \leq \theta \leq 1$$

Slide 7

- Then

$$p(\theta|y,n,M) = \frac{p(y|\theta,n,M)}{p(y|n,M)}$$

Justification for uniform prior

- Prior has to be $p(\theta|M) = 1$ if
  - prior predictive distribution (predictions before observations) is uniform

$$p(y|n) = \frac{1}{n+1}, \quad y = 0,\ldots,n$$

    - it seems that Bayes used this justification
    - nice justification, because it can be derived only from the observed quantities y and n

Slide 8

  - if all values of θ are equally probable
    - it seems that Laplace used this justification directly for θ using "indifference principle"

Binomial: posterior of θ

- Then

$$p(\theta|y,n,M) = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}d\theta} = \frac{1}{Z}\theta^y(1-\theta)^{n-y}$$

- Normalization

$$Z = \int_0^1 \theta^y(1-\theta)^{n-y}d\theta$$

Slide 9

Distributions and normalization

- Following

$$p(\theta|y,n,M) = \frac{1}{Z}\theta^y(1-\theta)^{n-y}$$

  is often replaced with this

$$p(\theta|y,n,M) \propto \theta^y(1-\theta)^{n-y}$$

Slide 10

- Unnormalized distributions are commonly used
  - normalization is computed at the final phase
  - or we can use, e.g., Monte Carlo methods to sample directly from the unnormalized distribution
- Terms
  - if $\int \pi(\theta)d\theta = \infty,\ \pi(\theta)$ improper
  - if $\int q(\theta)d\theta = Z \neq 1,\ q(\theta)$ unnormalized
  - if $\int p(\theta)d\theta = 1,\ p(\theta)$ is proper normalized

Binomial: posterior of θ

- Normalization

$$Z = p(y|n, M) = \int_0^1 \theta^y (1-\theta)^{n-y} d\theta = \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

- This is form of Beta function
  - when integrated over $(0, 1)$ the end result can be presented with Gamma functions
  - if y and n integers then $\Gamma(n) = (n-1)!$
  - for large n (and y) it may be better to compute $\log(\Gamma(\cdot))$

Slide 11

Binomial: posterior of θ

- Posterior

$$p(\theta|y, n, M) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \theta^y (1-\theta)^{n-y},$$

  is called Beta distribution

$$\theta|y, n \sim \text{Beta}(y+1, n-y+1)$$

Slide 12

Matlab demonstraatio: Beta-jakauma

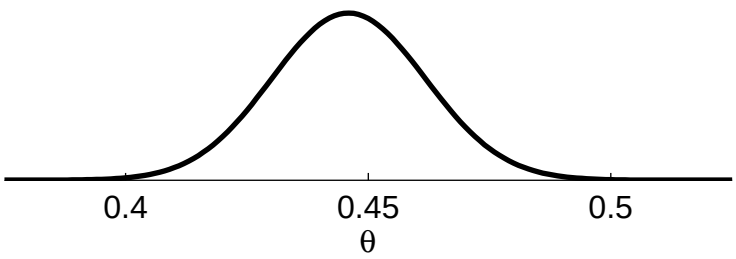
- disttool
  - $n = 2, y = 1 \rightarrow \text{Beta}(2, 2)$
  - $n = 5, y = 2 \rightarrow \text{Beta}(3, 4)$
  - $n = 20, y = 8 \rightarrow \text{Beta}(9, 13)$
  - $n = 100, y = 40 \rightarrow \text{Beta}(41, 61)$
  - $n = 980, y = 437 \rightarrow \text{Beta}(438, 544)$

Slide 13

Example: Probability of female birth

- 437 girls and 543 boys
  - probability of female birth?
  - compared to normal pr. 0.485?
- With uniform prior posterior is Beta$(438, 544)$

Slide 14



Presenting posterior distribution

- Posterior distribution contains all the current information about θ (given model and observations)
  - ideally report the whole distribution
  - often need to summarize, e.g., with location and width

Slide 15

Location of distribution

- Location summaries are called also point estimates
- Mean is posterior expectation $E(\theta|\cdot)$
- Median has equal amount of probability mass above and below
- Mode is a most probable value

Slide 16

- We could use other point estimates based on decision analysis using application specific cost function

Width of distribution

- Standard deviation describes width of the Gaussian, and thus can be used describe nearly Gaussian distributions
  - $\mathrm{Sd}(\theta|\cdot)$ is square root of $Var(\theta|\cdot)$

Slide 17

Example: pr. of female birth

- 437 girls and 543 boys
  - probability of female birth?
- With uniform prior posterior is Beta$(438, 544)$
- $\mathrm{Beta}(\alpha, \beta)$-distribution

Slide 18

$$\mathrm{E}(\theta) = \frac{\alpha}{\alpha + \beta} \approx 0.446$$

$$\mathrm{Sd}(\theta) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \approx 0.016$$

Posterior / credible interval

- Interval describes both location and width
- Posterior interval is also called
  - credible interval
  - Bayesian confidence interval
- Posterior interval contains specific (e.g. 95%) amount of the probability mass
  - not uniquely defined

Slide 19

- Most common choices
  - central posterior interval
  - highest posterior density (HPD) interval – shortest possible interval

Central posterior interval

- Central posterior interval
  - easy to compute
  - not good summary if mode in the edge of parameter space
  - not good summary if multimodal distribution
  - does not generalize for multidimensional distributions
  - invariant to one-to-one transformations

Slide 20

Highest posterior density (HPD) interval

- Highest posterior density (HPD) interval
  - almost as easy to compute as central
  - generalizes for multidimensional distributions
  - not invariant to one-to-one transformations

Slide 21

Probabilities

- Probabilities, Bayesian p-values

$$p(\theta \in A|y, M) = \int_{\theta \in A} p(\theta|y, M)d\theta$$

  - in one dimension

$$p(a < \theta < b) = \int_a^b p(\theta|y, n, M)$$

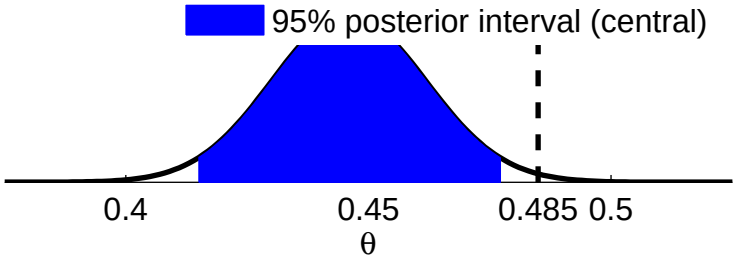
Slide 22

Example: pr. of female birth

- 437 girls and 543 boys
- With uniform prior posterior is Beta$(438, 544)$
  - mean 0.446 and std 0.016
  - 95% central interval $[0.415, 0.477]$
  - $p(\theta < 0.485) = 0.99$

Slide 23

- Matlab-demo: esim2_1.m



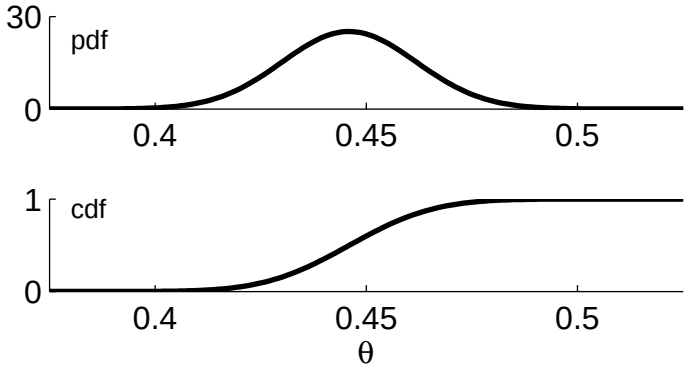
Cumulative density function (CDF)

- Cumulative density function (CDF)

$$p(\theta \leq a|\cdot) = \int_{-\infty}^{a} p(\theta|\cdot)p\theta$$

  - only for one-dimensional
  - Matlab has cdf-functions for most common densities

Slide 24



Summarizing posterior-distributions: computation

- Mean, median and std for many common distributionsin analytic form (appendix A)
- CDFs can be used easily to compute quantiles, intervals and probabilities
  - CDFs not necessarily in easy form
- Eg. integral $\int_0^x \theta^y(1-\theta)^{n-y}d\theta$ on muotoa incomplete Beta function

Slide 25

  - problematic for Bayes
  - nowadays several series and continued fraction approximations known
  - Matlab's betacdf-function uses continued fraction with and backup option
  - Laplace approximated using normal distribution

Example: pr. of female birth

- In Paris 1745–1770 241945 girls and 251527 boys
- Laplace approximated $p(\theta \geq 0.5|y,n,M) = \int_{0.5}^1 p(\theta|y,n,M)d\theta$
- Laplace used normal approximation (ch 4) and new how to calculate cdf for normal distribution

$$p(\theta \geq 0.5|y,n,M) \approx 1.15 \times 10^{-42}$$

Slide 26

- Laplace wrote that he is *'morally certain'*, that $\theta < 0.5$

Summarizing posterior-distributions: computation

- In general case numerical approximations
  - eg. quadrature integration
  - eg. Monte Carlo integration
    - approximate expectation using samples ($\theta^{(t)}$)

$$E(g(\theta)) \approx \frac{1}{N} \sum_{t=1}^{T} g(\theta^{(t)})$$

Slide 27

Problematic distributions

- Multimodal
- High-dimensional

Slide 28

About priors

- Prior should be nonzero for all slightly possible parameter values
  - if prior is 0, then posterior is 0
  - if prior is nonzero and there is a lot of data, data can swamp the prior

Slide 29

Conjugate priors

- Definition

  $$\text{if } p(\cdot|y) \in P \text{ for all } p(y|\cdot) \in F \text{ and } p(\cdot) \in P$$

  where P is a class of prior distributions and F is a class of sampling distributions

  this is too broad definition if P is the class of all distributions

- For *natural* conjugate priors P is the set off all densities having same functional form

Slide 30

Beta prior for Binomial

- Prior

$$\text{Beta}(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

$$\begin{aligned} p(\theta|y,n,M) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= \text{Beta}(\theta|\alpha+y,\beta+n-y) \end{aligned}$$

Slide 31

- $(\alpha-1)$ and $(\beta-1)$ virtual prior observations

Conjugate priors

- Conveniences of conjugate priors
  - interpretation
  - closed form posterior
  - computationally convenient
  - important building blocks for hierarchical models
  - usage expanded with mixtures

Slide 32

- Non-conjugate priors
  - computation more difficult, but not impossible
  - no need to make trade-of in presentation of the prior knowledge

Example of prior effect

- 437 girls and 543 boys
  - probability of female birth?
- With uniform prior ($\alpha = 1, \beta = 1$) posterior is Beta$(438, 544)$
- Test prior with mean 0.485 and "number of prior observations" 20 or 200
  - Matlab-demo: esim2_2.m

Slide 33

Example of Monte Carlo -computation

- 437 girls and 543 boys
  - whait we want to compute $\phi = (1 - \theta)/\theta$
  - $p(\phi|y, n, M) = ?$
- Easy to sample
  - first pick samples $\theta^{(t)}$ from $p(\theta|y, n, M)$
  - then $\phi^{(t)} = (1 - \theta^{(t)})/\theta^{(t)}$
  - $\phi^{(t)}$ are now samples from $p(\phi|y, n, M)$
  - histogram, quantiles and intervals easy to estimate from samples
  - Matlab-demo: esim2_3.m

Slide 34

Example of non-conjugate prior

- 437 girls and 543 boys
- Non-conjugate prior
  - posterior not in easy form
  - computation easy with Monte Carlo
  - eg. grid-sampling for one-dimensional

Slide 35

- Matlab-demo: esim2_4.m
  - this is also inverse-cdf demo

Binomial model prediction

- Uniform-prior

$$\begin{aligned} p(\tilde{y}=1|y,n,M) &= \int_0^1 p(\tilde{y}=1|\theta,y,n,M)p(\theta|y,n,M)d\theta \\ &= \int_0^1 \theta p(\theta|y,n,M)d\theta \\ &= \mathrm{E}[\theta] = \frac{y+1}{n+2} \end{aligned}$$

Slide 36

- Extreme cases

$$p(\tilde{y}=1|y=0,n,M) = \frac{1}{n+2}$$
$$p(\tilde{y}=1|y=n,n,M) = \frac{n+1}{n+2}$$

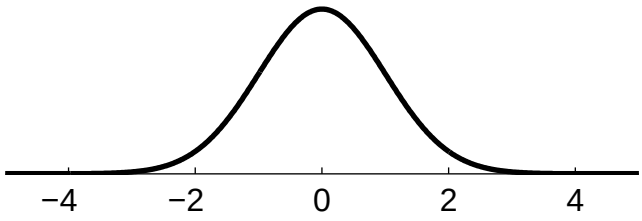
- Compare to maximum likelihood

Gaussian

- Commonly used alone and as part of hierarchical models
- Observation y gets real values
- Parameters mean θ and variance σ^2
  (we first assume σ^2 known)

Slide 37

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$
$$y \sim \mathrm{N}(\theta, \sigma^2)$$



Justifications used for Gaussian

- Central limit theorem
- Exchangeability and ball symmetry
- Computational conveniency

Slide 38

Central limit theorem*

- Tells how sum (avergare) of random variables behaves when $n \to \infty$
- De Moivre, Laplace, Gauss, Chebysev, Liapounov, Markov, et al.
- With certain conditions distribution of sum (or average) goes to Gaussian when $n \to \infty$

Slide 39

- Eg. if different sources of uncertainty, we may asume that sum is near Gaussian
- Problems
  - does not always hold, eg. Cauchy-jakauma
  - may need big n,
    eg. with Binomial, if θ near 0 or 1
  - does not hold if variance of one of the variables dominates
- Eg:
  `http://noppa5.pc.helsinki.fi/koe/flash/clt/clt2.html`

About central limit theorem

- Sensible if assumed that uncertainty is sum of several exchangeable or independent quantities
  - having similar scale
  - having similar thicknes of tails

Slide 40

Computational conveniency

- Negative log-likelihood has nice form

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$
$$-\log p(y|\theta) = \alpha(y-\theta)^2 + C$$

Slide 41

- Minimizing neg-log-likelihood is equal to smallest square method by (Gauss)
- Analysis for linear models with Gaussian likelihood can be computed with simple matrix calculus

Gaussian

- Used in many models, eg:
  - as alone
  - part of hierarchical models
  - part of mixture models
  - part of scale mixture presentation of t-distribution
    - t-distribution is good more robust alternative to Gaussian

Slide 42

Gaussian and conjugate prior

- Assume, that σ known

Likelihood $$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

Prior $$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

Slide 43

Posterior $$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$

Gaussian and conjugate prior

- Posterior (ex 2.14a)

$$\begin{aligned} p(\theta|y) &\propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right) \\ &\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right) \end{aligned}$$

Slide 44

$$\theta|y \sim \mathrm{N}(\mu_1, \tau_1^2), \quad \text{missä} \quad \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}} \quad \text{ja} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2}+\frac{1}{\sigma^2}$$

- 1/variance = precision

Gaussian - example

- Population IQ: $\theta \sim \mathrm{N}(100, 15^2)$ and observation: $y|\theta \sim \mathrm{N}(\theta, 10^2)$
  estimate person's IQ given observation y

$$\mathrm{E}(\theta|y) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} y + \frac{\sigma^2}{\tau_0^2 + \sigma^2} \mu_0$$

$$\mathrm{Std}(\theta|y) = \left(\frac{1}{\tau_0^2} + \frac{1}{\sigma^2} \right)^{-1/2}$$

Slide 45

$$\tau_0 = 15, \sigma = 10: \quad \mathrm{E}(\theta|y) \approx 0.7y + 30 \quad \text{ja} \quad \mathrm{sd}(\theta|y) \approx 8$$

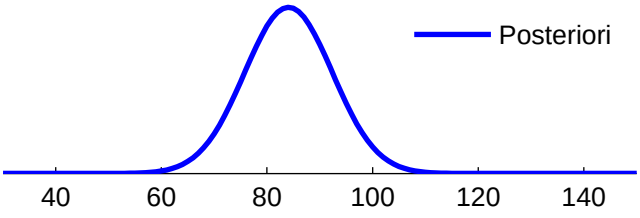
- compare to maximum likelihood

$$\mathrm{E}(\theta|y) = y \quad \text{ja} \quad \mathrm{sd}(\theta|y) = 10$$

Gaussian - example

- Test $y = 77$ (esim3_1.m)
  $\mathrm{E}(\theta|y) \approx 84$
  $\mathrm{sd}(\theta|y) \approx 8$
  $p(\theta > 100|y) \approx 0.03$

Slide 46



Gaussian

- Posterior predictive distribution

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$
$$p(\tilde{y}|y) \propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right)\exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)d\theta$$

Slide 47

$$\tilde{y}|y \sim \mathrm{N}(\mu_1, \sigma^2+\tau_1^2)$$

Gaussian - example

- Population IQ: $\theta \sim \mathrm{N}(100, 15^2)$ and observation: $y|\theta \sim \mathrm{N}(\theta, 10^2)$ distribution for the next test result $\tilde{y}$ of the same person given first result y

$$\mathrm{E}(\tilde{y}|y) = \mu_1 \approx 0.7y + 30$$
$$\mathrm{Std}(\tilde{y}|y) = (\sigma^2+\tau_1^2)^{1/2} \approx 13$$

Slide 48

Gaussian - example

- Test result $y = 77$ (esim3_1.m)

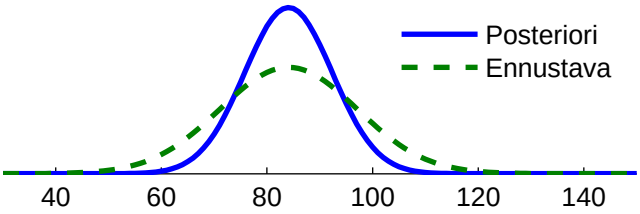
  $\mathrm{E}(\theta|y) \approx 84$

  $\mathrm{sd}(\theta|y) \approx 8$

  $\mathrm{E}(\tilde{y}|y) \approx 84$

  $\mathrm{sd}(\tilde{y}|y) \approx 13$

Slide 49



Gaussian - several observations

- Several observations $y = (y_1, \ldots, y_n)$ and assume that we can proceed as if they were independent identically distributed

Slide 50

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &= p(\theta)\prod_{i=1}^{n} p(y_i|\theta) \\ &= \mathrm{N}(\theta|\mu_n, \tau_n^2) \end{aligned}$$

where $\mu_n = \dfrac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$ ja $\dfrac{1}{\tau_n^2} = \dfrac{1}{\tau_0^2} + \dfrac{n}{\sigma^2}$

where $\bar{y} = \frac{1}{n}\sum_i y_i$

- see ex 2.14b

Sufficient statistic

- $t(y)$ is sufficient statistic, if likelihood of θ depends on data y only through $t(y)$
- Examples
  - Binomial: $t(y_1, \ldots, y_n) = (\sum_i y_i, n)$
  - Gaussian with known variance: $t(y_1, \ldots, y_n) = (\frac{1}{n}\sum_i y_i, n) = (\bar{y}, n)$

Slide 51

Gaussian - several observations

- Several observations $y = (y_1, \ldots, y_n)$

$$p(\theta|y) = \mathrm{N}(\theta|\mu_n, \tau_n^2)$$

$$\text{where} \quad \mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{ja} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

Slide 52

- If $\tau_0^2 = \sigma^2$ prior corresponds to one prior observation with mean μ_0
- If $\tau_0 \to \infty$ when n fixed
  or if $n \to \infty$ when τ_0 fixed

$$p(\theta|y) \approx \mathrm{N}(\theta|\bar{y}, \sigma^2/n)$$

Gaussian - known mean

- Likelihood

$$p(y|\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}v\right)$$

$$\text{where} \quad v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$$

Slide 53

- Conjugate prior is inverse-gamma

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right)$$

- Somewhat more convenient parameterization $\text{Inv-gamma}(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2}s^2)$

$$p(\sigma^2) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^{\nu} (\sigma^2)^{-(\nu/2+1)} \exp(-\nu s^2/(2\sigma^2))$$
$$\sigma^2 \sim \text{Inv-}\chi^2(\nu, s^2)$$

Gaussian - known mean

- With parameterization

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

the posterior is

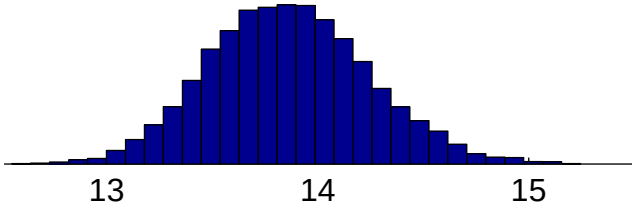
$$\sigma^2|y \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + nv}{\nu_0 + n}\right)$$

Slide 54

Gaussian - known mean - example

- Football data, model $N(0, \sigma^2)$
- $\nu_0 = 0$ corresponds $p(\sigma^2) \propto \sigma^{-2}$ (improper)
- Posterior is anyway proper, $\sigma^2|d \sim \text{Inv-}\chi^2(n, v)$, $n = 672$ ja $v = 13.85^2$

Slide 55



Poisson

- Model for number of events which are exchangeable in time
  - e.g. independent events, having each moment same probability of happening
- Used e.g. in epidemiology to estimate probabilities to get some disease
- Likelihood

Slide 56

$$p(y|\theta) \propto \theta^{t(y)} e^{-n\theta}, \quad \text{where} \quad t(y) = \sum_{i=1}^{n} y_i$$

- Conjugate prior is gamma and posterior is

$$\theta|y \sim \text{Gamma}(\alpha + n\bar{y}, \beta + n)$$

Poisson - example

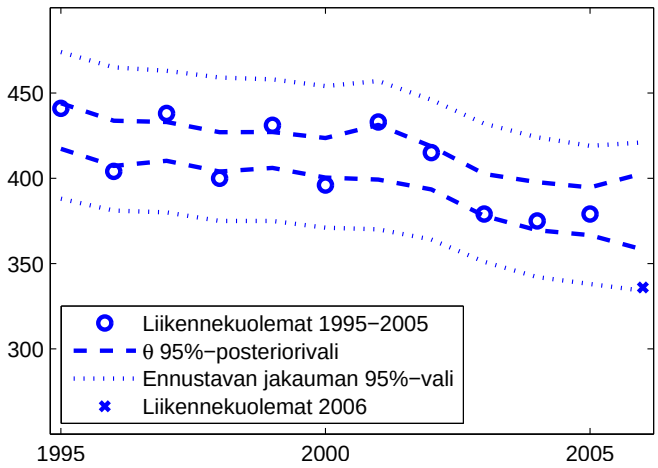
- According to Liikenneturva in recent years about 400 dies in traffic each year
- In 2006 336 died in traffic
- Is year 2006 exceptional?
  - $p(y_{2006} \leq 336 | \theta = 400) \approx 10^{-4}$
  - $p(y_{2006} \leq 336 | y_{1995,\ldots,2005}, \text{constant risk}) \approx 4 \times 10^{-4}$

Slide 57

Poisson - example

- Traffic safety changing slowly?
  - risk going down?
  - time series model (log risk has Gaussian process prior)
    $p(y_{2006} \leq 336 | y_{1995,\ldots,2005}, \text{changing risk}) \approx 0.03$

Slide 58

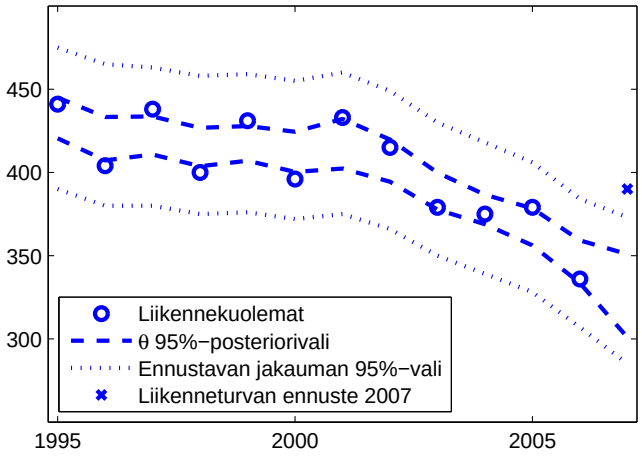


Poisson - example

Slide 59

- Based on beginning of 2007 Liikenneturva predicted 2007 390 dies
  - add observation from year 2006

    $p(y_{2007} \leq 390 | y_{1995,\ldots,2006}, \text{changing risk}) \approx 0.99$



Exponential

Slide 60

- Model for waiting time of events which are exchangeable in time
- E.g. survival times
- Likelihood

$$p(y|\theta) = \theta \exp(-y\theta), y > 0$$

- Conjugate prior is gamma and posterior is

$$\text{Gamma}(\theta|\alpha + n, \beta + n\bar{y})$$

Conjugate priors and sufficient statistics

- Generally, only distributions in exponetial family have conjugate prior
- Only distributions in exponetial family, have sufficient statistics, (execpt some irregular cases, like uniform)
- Exponential family has form

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)}$$

Slide 61

- So far distributions mentioned have belonged to this family (expect uniform)

Cauchy

- Likelihood $p(y_i|\theta) = 1/(1 + (y_i - \theta)^2)$
- Infinite variance, i.e has very long tails
- For example
  - ratio of two zero mean Gaussian quantities X/Y
  - resonance models in physics
  - spectroscopy

Slide 62

- Cauchy (of half-Cauchy) is also used as robust prior

Non-informative priors

- Limits of conjugate priors
- Indifference
- Jeffreys' prior
- Reference priors

Slide 63

- Hierarchical priors

Limits of conjugate prior

- Conjugate prior is non-informative if number of virtual prior samples is zero

Slide 64

Transformation of variables

- Example
  - If $p(\sigma) = 1/\sigma$ then $p(\log(\sigma))$

$$
\begin{aligned}
p(\log\sigma) &= |J|p(\sigma) && \text{(book s. 24)} \\
&= \frac{d\sigma}{d(\log\sigma)}\frac{1}{\sigma} \\
&= \sigma\frac{1}{\sigma} \\
&= 1 && \text{(improper)}
\end{aligned}
$$

Slide 65

Jeffreys' prior (indifferencen yleistys)

- Prior is invariant to transformation of variable
- Fisher's information matrix is $I(\theta)$, where $I(\theta)_{ij} = E\left(-\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\right)$
- Jeffreys' prior is

$$p(\theta) \propto \det(I(\theta))^{1/2}$$

Slide 66

- Problematic for multiparameter models
- Usually location, scale and mixing parameters handled separately
- E.g.:
  $y \sim \text{Bin}(n,\theta): \quad p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$
  $y \sim N(\mu,\sigma^2): \quad p(\mu,\sigma^2) \propto 1/\sigma^2$

Reference prior (Bernardo ja Berger-Bernardo)*

- Generalizes Jeffreys's prior
  - same in simple cases
- Information theoretic definition
- Produces better results for multiparameter models
  - prior still depends on focus, ordering and grouping

Slide 67

Dangers of non-informative priors

- Vague priors may be sensitive to parameterization
- Some of the methods produce improper prior, and the it must be checked whether posterior is proper
- Do not eliminate the need to think

Slide 68

Weakly informative priors

- Take into account even small amount of information available
  - usually with thick tails, being more robust if the used prior information in conflict with the data
  - e.g. vague t and Cauchy distributions

Slide 69

Hierarchical priors

- If you don't know suitable value for prior parameters, make it unknown haing it's own prior, etc.

Slide 70