Mathematical Statistics

Anna Janicka

Lecture XIV, 7.06.2021

Bayesian Statistics

Some concluding remarks

Plan for Today

1. Bayesian Statistics
   - a priori and a posteriori distributions
   - Bayesian estimation:
     - Maximum a posteriori probability (MAP)
     - Bayes Estimator
2. Caution!

Bayesian Statistics vs. traditional statistics

Frequentist: unknown parameters are given (fixed), observed data are random

Bayesian: observed data are given (fixed), **parameters are random**

Bayesian Statistics

Our knowledge about the unknown parameters is described by means of probability distributions, and additional knowledge may affect our description.

Knowledge:

- general
- specific

Example: coin toss

Bayesian Model

- $X_1, \ldots, X_n$ come from distribution P_θ, with density $f_\theta(\mathbf{x})$ – conditional density given a specific value of θ (likelihood function).
- $\mathscr{P}$ – family of probability distributions P_θ, indexed by the parameter $\theta \in \Theta$
- General knowledge: distribution Π over the parameter space Θ, given by $\pi(\theta)$ – the so-called **a priori/prior** distribution of θ,

$$\theta \sim \Pi$$

Bayesian Model – cont.

Additional knowledge (specific, contextual): based on observation. We have a joint distribution of observations and θ:

$$f(x_1, x_2, \dots, x_n, \theta) = f(x_1, x_2, \dots, x_n | \theta)\pi(\theta)$$

on this basis we can derive the conditional distribution of θ (given the observed data)

$$\pi(\theta | x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n | \theta)\pi(\theta)}{m(x_1, \dots, x_n)},$$

where

$$m(x_1, \dots, x_n) = \int_\Theta f(x_1, \dots, x_n | \theta)\pi(\theta) d\theta$$

is a marginal distribution for the obs.

Bayesian Model – a posteriori distribution

$\pi(\theta|x_1, \ldots, x_n)$ is called the **a posteriori/ posterior** distribution, denoted Π_x

The posterior distribution reflects all knowledge: general (initial) and specific (based on the observed data).

Grounds for Bayesian inference and modeling

Prior and posterior distributions: examples

1. Let $X_1, \ldots, X_n$ be IID r.v. from a 0-1 distr. with prob. of success θ, let

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

for $\theta \in (0,1)$

Beta(α,β) distr with mean $= \alpha/(\alpha+\beta)$

where $B(\alpha,\beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

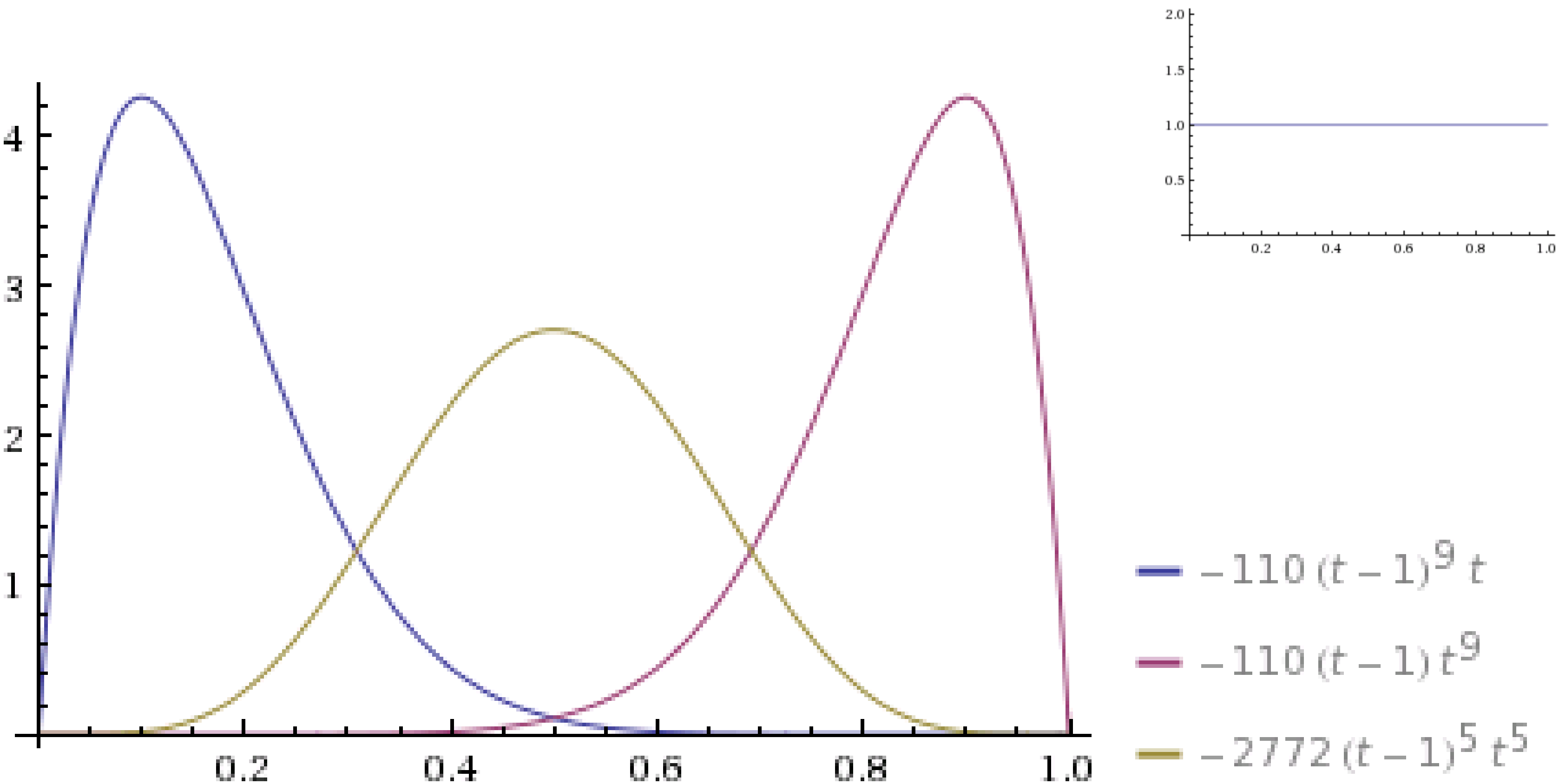
and $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}\exp(-u)du = (\alpha-1)\Gamma(\alpha-1)$

then the posterior distribution:

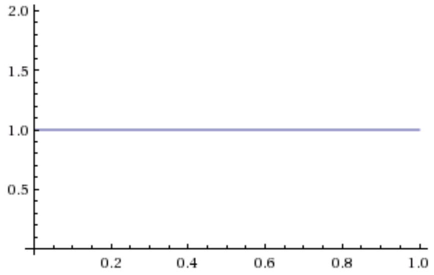
$$\mathrm{Beta}\left(\sum_{i=1}^n x_i + \alpha, n - \sum_{i=1}^n x_i + \beta\right)$$

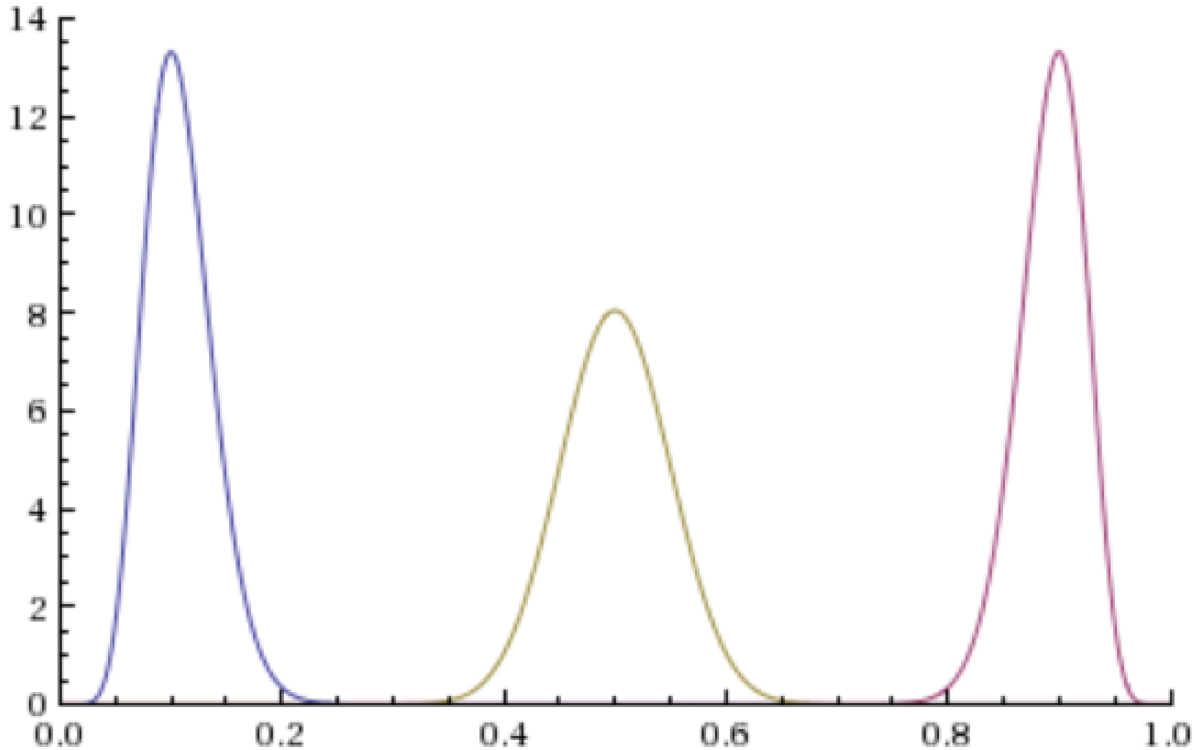
conjugate prior for Bernoulli distr.

For a Beta (1,1) prior and data: n=10 and 1, 5, 9 successes

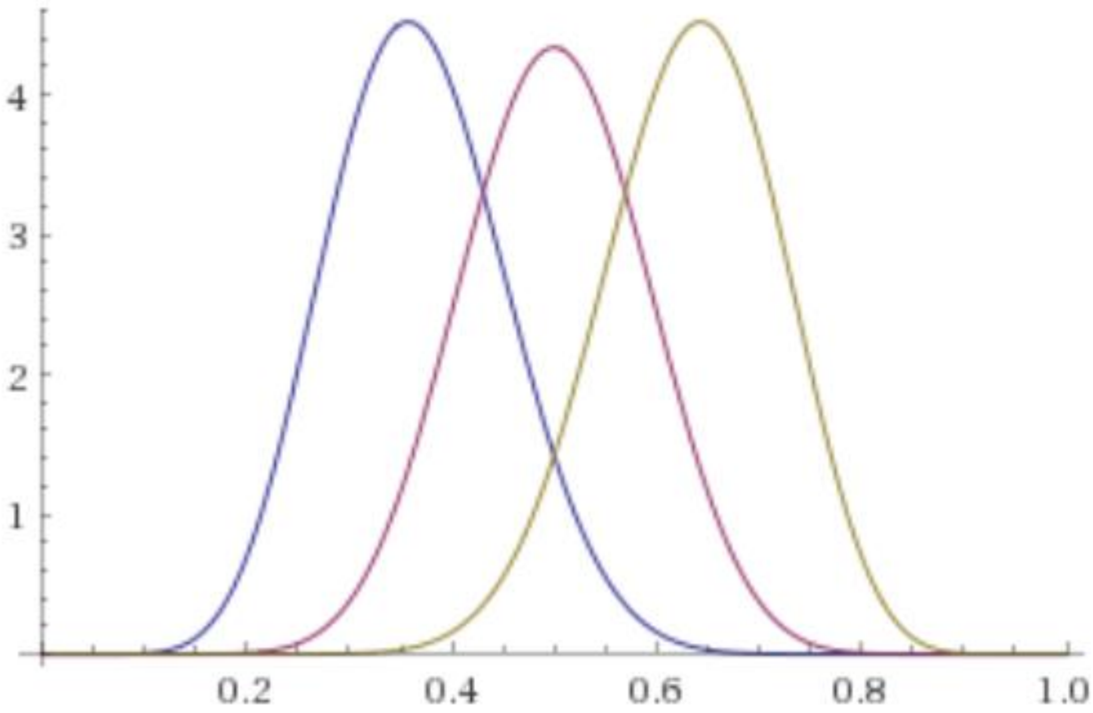


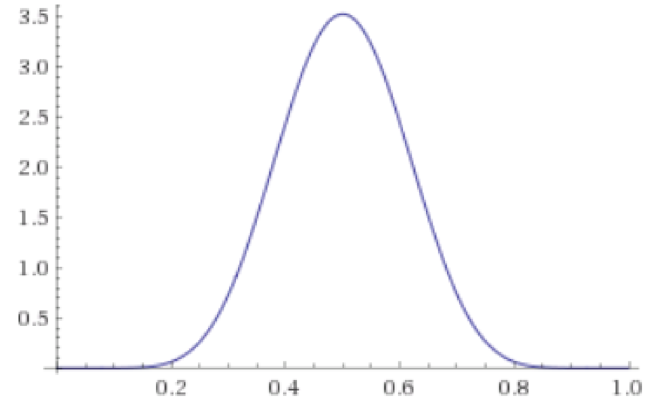
For a Beta (1,1) prior and data: n=100 and 10, 50, 90 successes



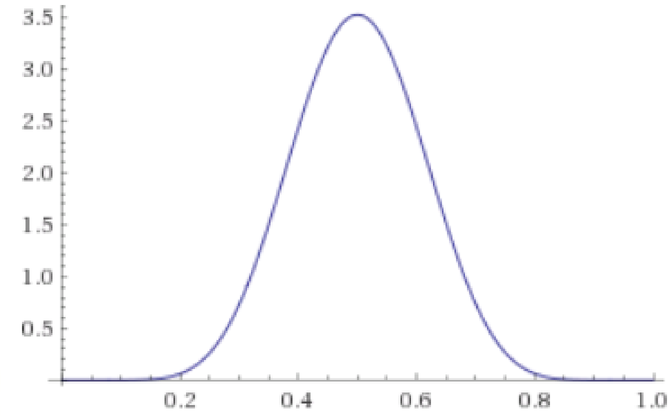


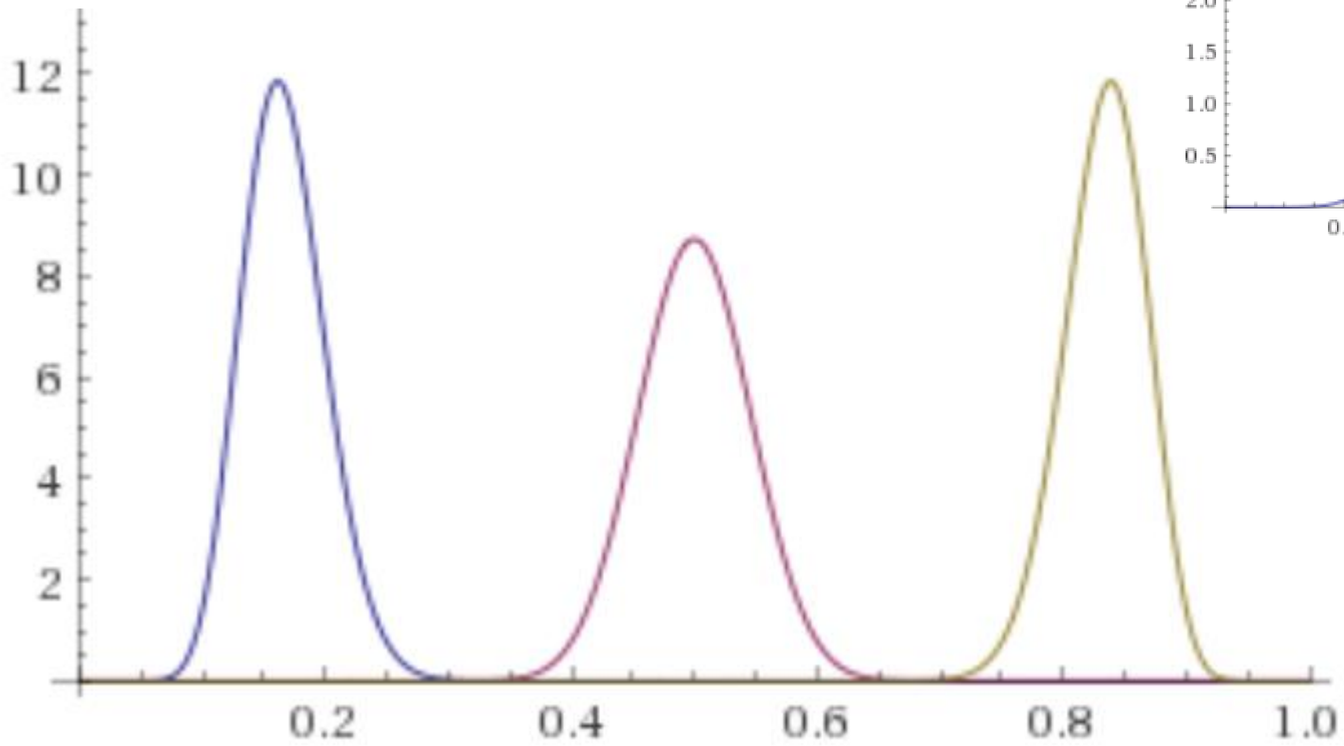
For a Beta (10,10) prior and data: n=10 and 1, 5, 9 successes



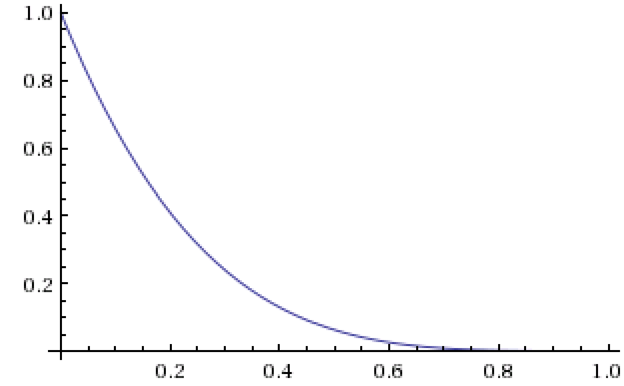


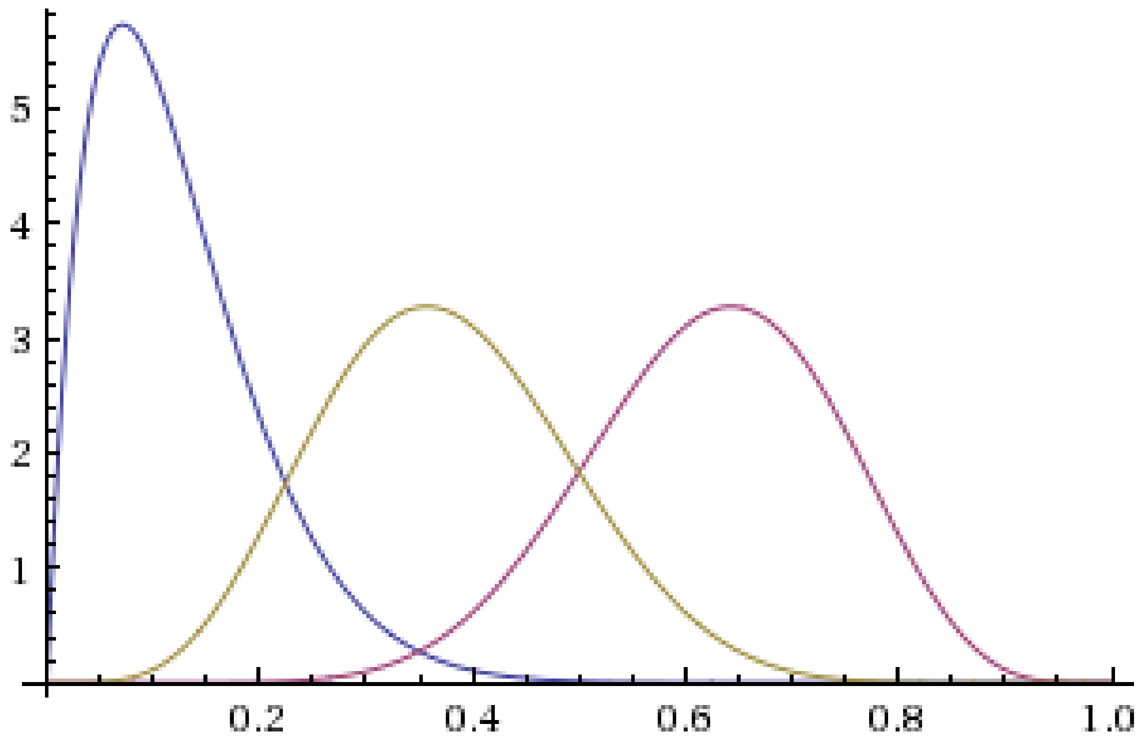
For a Beta (10,10) prior and data: n=100 and 10, 50, 90 successes



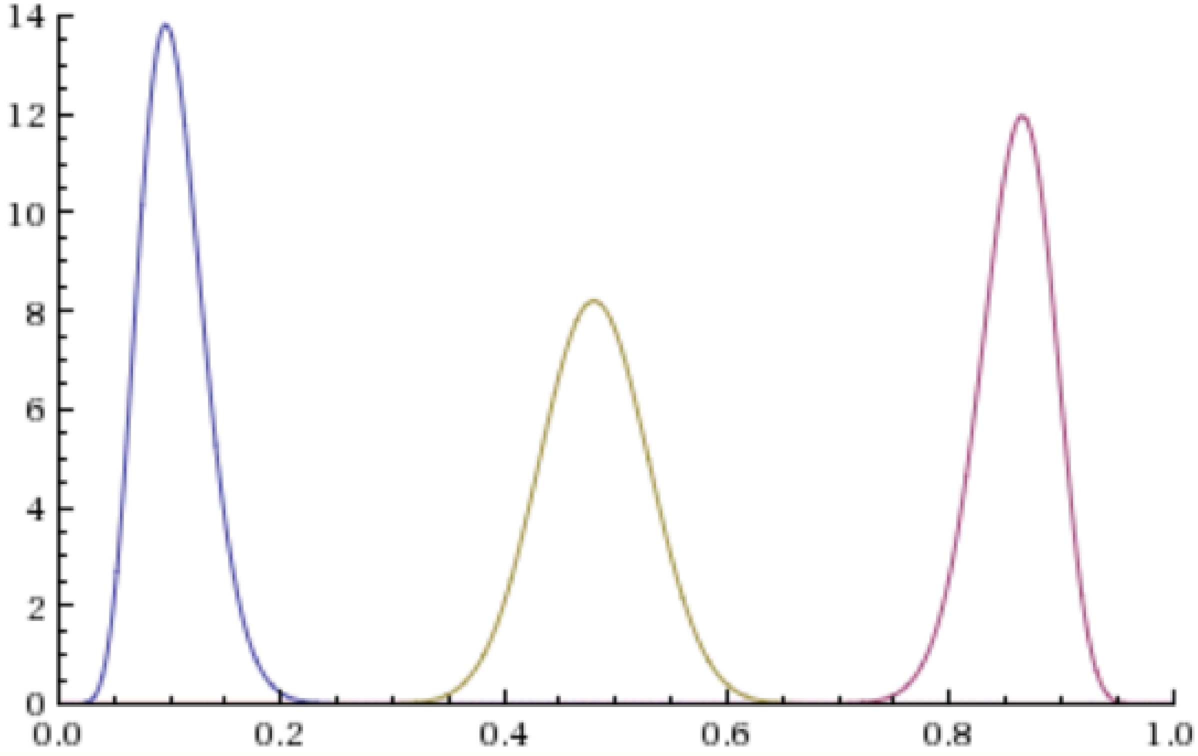


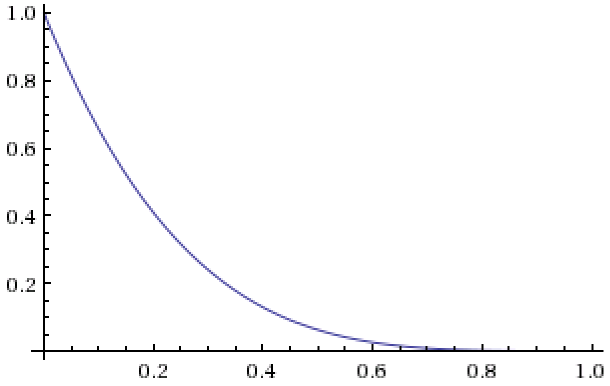
For a Beta (1,5) prior and data: n=10 and 1, 5, 9 successes





For a Beta (1,5) prior and data: n=100 and 10, 50, 90 successes





Prior and posterior distributions: examples (2)

2. Let X_1, ..., X_n be IID r.v. from $N(\theta, \sigma^2)$, and σ^2 known; $\theta \sim N(m, \tau^2)$ for m, τ known.

Then the posterior distribution for θ:

$$N\left(\frac{n\frac{1}{\sigma^2}\bar{X}+\frac{1}{\tau^2}m}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}},\frac{1}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}\right)$$

conjugate prior for a normal distr.

Bayesian Statistics

Based on the Bayes approach, we can

- find estimates
- find an equivalent of confidence intervals
- verify hypotheses

- make predictions

Bayesian Most Probabale (BMP) / Maximum a posteriori Probability (MAP) estimate

Similar to ML estimation: the argument which maximizes the posterior distribution:

$$\pi(\hat{\theta}_{BMP}|x_1,\ldots,x_n) = \max_\theta \pi\,(\theta|x_1,\ldots,x_n)$$

i.e.

$$BMP(\theta) = \hat{\theta}_{BMP} = \operatorname{argmax}_\theta \pi\,(\theta|x_1,\ldots,x_n)$$

BMP: examples

1. Let $X_1, ..., X_n$ be IID r.v. from a Bernoulli distr. with prob. of success θ; for $\theta \in (0,1)$

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

We know the posterior distribution:

$$\mathrm{Beta}(\sum_{i=1}^{n} x_i + \alpha, n - \sum_{i=1}^{n} x_i + \beta)$$

Beta(α,β) distr; the mode of this distr = (α-1)/(α+ β-2) for α>1, β>1

we have max for

$$BMP(\theta) = \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{n + \beta + \alpha - 2}$$

i.e. for 5 successes in 10 trials for a prior U(0,1) (i.e. Beta(1,1) distr.), we have $BMP(\theta)$=5/10 = ½

and for 9 successes in 10 trials for the same prior distr., we have $BMP(\theta)$=9/10

BMP: examples (2)

2. Let $X_1, ..., X_n$ be IID r.v. from $N(\theta, \sigma^2)$, with σ^2 known; $\theta \sim N(m, \tau^2)$ for m, τ known.

Then the posterior distr. for θ:

$$N\left(\frac{n\frac{1}{\sigma^2}\bar{X}+\frac{1}{\tau^2}m}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}\right)$$

so

$$BMP(\theta) = \frac{n\frac{1}{\sigma^2}\bar{X}+\frac{1}{\tau^2}m}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}$$

i.e. if we have a sample of 5 obs 1.2; 1.7 ; 1.9 ; 2.1; 3.1 from distr. N(θ, 4) and the prior distr is $\theta \sim N(1, 1)$, then

$$BMP(\theta) = (5/4 * 2 + 1)/(5/4 + 1) = 14/9 \approx 1.56$$

and if the prior distr were $\theta \sim N(3, 1)$, then

$$BMP(\theta) = (5/4 * 2 + 1*3)/(5/4 + 1) = 22/9 \approx 2.44$$

Bayes Estimator

An estimation rule which minimizes the posterior expected value of a loss function

$L(\theta, a)$ – **loss function,** depends on the true value of θ and the decision a.

e.g. if we want to estimate $g(\theta)$:

$L(\theta, a) = (g(\theta) - a)^2$ – quadratic loss function

$L(\theta, a) = |g(\theta) - a|$ – module loss function

Bayes Estimator – cont.

We can also define the **accuracy of an estimate** for a given loss function :

$$acc(\Pi, \hat{g}(x)) = E(L(\theta, \hat{g}(x))|X = x) = \int_{\Theta} L(\theta, \hat{g}(x))\pi(\theta|x)d\theta$$

(the average loss of the estimator for a given prior distribution and data, i.e. for a specific posterior distribution)

Bayes Estimator – cont. (2)

The **Bayes Estimator** for a given loss function $L(\theta, a)$ is $\hat{g}_B$ such that

$$\forall x \quad acc(\Pi, \hat{g}_B(x)) = \min_a a\, cc(\Pi, a)$$

For a quadratic loss function $(\theta - \mathrm{a})^2$:

$$\hat{\theta}_B = E(\theta|X = x) = E(\Pi_x)$$

For a module loss function $|\theta - \mathrm{a}|^2$:

$$\hat{\theta}_B = Med(\Pi_x)$$

more generally: $E(g(\theta)|x)$

Bayes Estimator: Example (1)

1. Let $X_1, \ldots, X_n$ be IID r.v. from a Bernoulli distr. with prob. of success θ; for $\theta \in (0,1)$

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

Beta(α,β) distr with mean = $\alpha/(\alpha+\beta)$

We know the posterior distribution:

$$\mathrm{Beta}\left(\sum_{i=1}^{n} x_i + \alpha, n - \sum_{i=1}^{n} x_i + \beta\right)$$

so the Bayes Estimator is

$$\hat{\theta}_B = \frac{\sum_{i=1}^{n} x_i + \alpha}{n + \beta + \alpha}$$

i.e. for 5 successes in 10 trials for a prior U(0,1) (i.e. Beta(1,1) distr.), we have $\hat{\theta}_B$ =6/12 = ½

and for 9 successes in 10 trials for the same prior distr., we have $\hat{\theta}_B$ =10/12 = 5/6

BMP: examples

1. Let $X_1, \ldots, X_n$ be IID r.v. from a Bernoulli distr. with prob. of success θ; for $\theta \in (0,1)$

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

Beta(α,β) distr; the mode of this distr = (α-1)/(α+ β-2) for α>1, β>1

We know the poster distribution:

$$\mathrm{Beta}(\sum_{i=1}^{n} x_i + \alpha, n - \sum_{i=1}^{n} x_i + \beta)$$

we have max for

$$BMP(\theta) = \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{n + \beta + \alpha - 2}$$

i.e. for 5 successes in 10 trials for a prior U(0,1) (i.e. Beta(1,1) distr.), we have *BMP*(θ)=5/10 = ½

and for 9 successes in 10 trials for the same prior distr., we have *BMP*(θ)=9/10

Bayes Estimator: examples (2)

2. Let $X_1, \ldots, X_n$ be IID r.v. from $N(\theta, \sigma^2)$, with σ^2 known; $\theta \sim N(m, \tau^2)$ for m, τ known.

Then the a posteriori distr for θ:

$$N\left(\frac{n\frac{1}{\sigma^2}\bar{X}+\frac{1}{\tau^2}m}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}\right)$$

so

$$\hat{\theta}_B = \frac{n\frac{1}{\sigma^2}\bar{X}+\frac{1}{\tau^2}m}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}$$

i.e. if we have sa sample of 5 obs 1.2; 1.7 ; 1.9 ; 2.1; 3.1 from distr. N(θ, 4) and the prior distr is $\theta \sim N(1, 1)$, then

$\hat{\theta}_B$ = (5 /4 * 2 + 1)/(5/4 + 1) = 14/9 ≈ 1.56

and if the prior distr were $\theta \sim N(3, 1)$, then

$\hat{\theta}_B$ = (5 /4 * 2 + 1*3)/(5/4 + 1) = 22/9 ≈ 2.44

BMP: examples (2)

2. Let $X_1, \ldots, X_n$ be IID r.v. from $N(\theta, \sigma^2)$, with σ^2 known; $\theta \sim N(m, \tau^2)$ for m, τ known.

Then the a posteriori distr for θ: $N\left(\frac{n\frac{1}{\sigma^2}\bar{X}+\frac{1}{\tau^2}m}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}\right)$

so

$$BMP(\theta) = \frac{n\frac{1}{\sigma^2}\bar{X}+\frac{1}{\tau^2}m}{n\frac{1}{\sigma^2}+\frac{1}{\tau^2}}$$

i.e. if we have sa sample of 5 obs 1.2; 1.7 ; 1.9 ; 2.1; 3.1 from distr. N(θ, 4) and the prior distr is $\theta \sim N(1, 1)$, then

$$BMP(\theta) = (5/4 * 2 + 1)/(5/4 + 1) = 14/9 \approx 1.56$$

and if the prior distr were $\theta \sim N(3, 1)$, then

$$BMP(\theta) = (5/4 * 2 + 1*3)/(5/4 + 1) = 22/9 \approx 2.44$$

Caution!

1. Tests should be designed BEFORE we start examining the data
2. The only way to increase power and improve significance level simultaneously is by collecting more observations (frequently not possible if we work on existing data).
3. Significant p-value does not mean effect is important/sizeable.
4. P-values in repeated samples

P-values in repeated samples

We examine if a new training has effect. The null hypothesis is that the training has no effect, and the alternative hypothesis is that it has effect. We use a 5% significance level for the test.

- A randomly selected school has completed this training, and after completion the statistical test returns a *P*-value equal to 4%.
- 25 different schools have completed this training. At one of the schools the test returned a *P*-value of 4%.