Mathematical Statistics

Anna Janicka

Lecture XIV, 27.05.2019

BAYESIAN STATISTICS

Plan for Today

1. Chi-squared tests – cont.
2. Bayesian Statistics
   - a priori and a posteriori distributions
   - Bayesian estimation:
     - Maximum a posteriori probability (MAP)
     - Bayes Estimator

Chi-squared goodness-of-fit test – reminder.

General form of the test:

$$\chi^2 = \sum \frac{(\text{observed value - expected value})^2}{\text{expected value}}$$

here: $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$ or $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i(\theta))^2}{np_i(\theta)}$

Theorem. If H_0 is true, for $n \to \infty$ the distribution of the χ^2 statistic converges to a chi-squared distr. with k-1 degrees of freedom $\chi^2(k\text{-}1)$ or to a chi-squared distr. with k-d-1 degrees of freedom $\chi^2(k\text{-}d\text{-}1)$ (depending on the dimension d of unknown parameter θ)

Chi-squared goodness-of-fit test – version for continuous distributions

Kolmogorov tests are better, but the chi-squared test may also be used

Model: $X_1, X_2, \ldots, X_n$ are an IID sample from a continuous distribution.

H_0: The distribution is given by F

H_1: $\neg H_0$ (i.e. the distribution is different)

It suffices to divide the range of values of the random variable into classes and count the observations. The expected values are known (result from F).Then: the chi-squared test.

Chi-squared goodness-of-fit test – practical notes

- The test should be used for large samples
- The expected counts can’t be too small (<5). If they are smaller, observations should be grouped.
- The classes in the „continuous” version may be chosen arbitrarily, but it is best if the theoretical probabilities are balanced.

Chi-squared test of independence

Model: (X_1, Y_1), ..., (X_n, Y_n) are an IID sample from a two-dimensional distribution with $r*s$ values (denoted by the set $\{1, ..., r\} \times \{1, ..., s\}$).

Let the theoretical distribution be

$$p_{ij} = P(X = i, Y = j) \quad i = 1,...,r \quad j = 1,...,s$$

Denote $$p_{i\bullet} = \sum_{j=1}^{s} p_{ij}, \quad p_{\bullet j} = \sum_{i=1}^{r} p_{ij}$$

We want to veryfy independence of X and Y:

H_0: $p_{ij} = p_{i\bullet} * p_{\bullet j} \quad i = 1,...,s, \quad j = 1,...,r$

H_1: $\neg H_0$

Chi-squared test of independence – cont.

The empirical distribution may be summarized by a table (so-called contingency table, or crosstab)

$i \setminus j$	1	2	...	s	$N_{i\bullet}$
1	N_{11}	N_{12}		N_{1s}	$N_{1\bullet}$
2	N_{21}	N_{22}		N_{2s}	$N_{2\bullet}$
...					
r	N_{r1}	N_{r2}		N_{rs}	$N_{r\bullet}$
$N_{\bullet j}$	$N_{\bullet 1}$	$N_{\bullet 2}$		$N_{\bullet s}$	n

Chi-squared test of independence – cont. (2)

This is a special case of a goodness-of-fit test with (r-1) + (s-1) parameters to be estimated:

The test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - N_{i\bullet} N_{\bullet j}/n)^2}{N_{i\bullet} N_{\bullet j}/n}$$

has a chi-squared distribution with (r-1)(s-1) degrees of freedom (if H_0 is true)

Chi-squared test of independence – example

We verify independence of political and musical preferences, for signif. level α =0.05

	Support X	Do not support X	Total
Listen to jazz	25	10	35
Listen to rock	20	20	40
Listen to hip-hop	15	10	25
Total	60	40	100

$$\chi^2 = \frac{(25-60*35/100)^2}{60*35/100} + \frac{(20-60*40/100)^2}{60*40/100} + \frac{(15-60*25/100)^2}{60*25/100}$$
$$+ \frac{(10-40*35/100)^2}{40*35/100} + \frac{(20-40*40/100)^2}{40*40/100} + \frac{(10-40*25/100)^2}{40*25/100} \approx 3.57$$

$$\chi^2_{1-0.05}((2-1)(3-1)) = \chi^2_{0.95}(2) \approx 5.99$$

→ **no grounds to reject H_0.**

Source: W. Niemiro

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).
- 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:

$$f(x_1, x_2, \ldots, x_n, \theta) = f(x_1, x_2, \ldots, x_n \mid \theta)\pi(\theta)$$

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

$$\pi(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{m(x_1, \ldots, x_n)},$$

where $m(x_1, \ldots, x_n) = \int_\Theta f(x_1, \ldots, x_n \mid \theta)\pi(\theta)d\theta$

is a marginal distribution for the obs.

Bayesian Model – a posteriori distribution

$\pi(\theta \mid x_1,\ldots,x_n)$ 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

A priori and a posteriori 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 $\theta \in (0,1)$

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

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

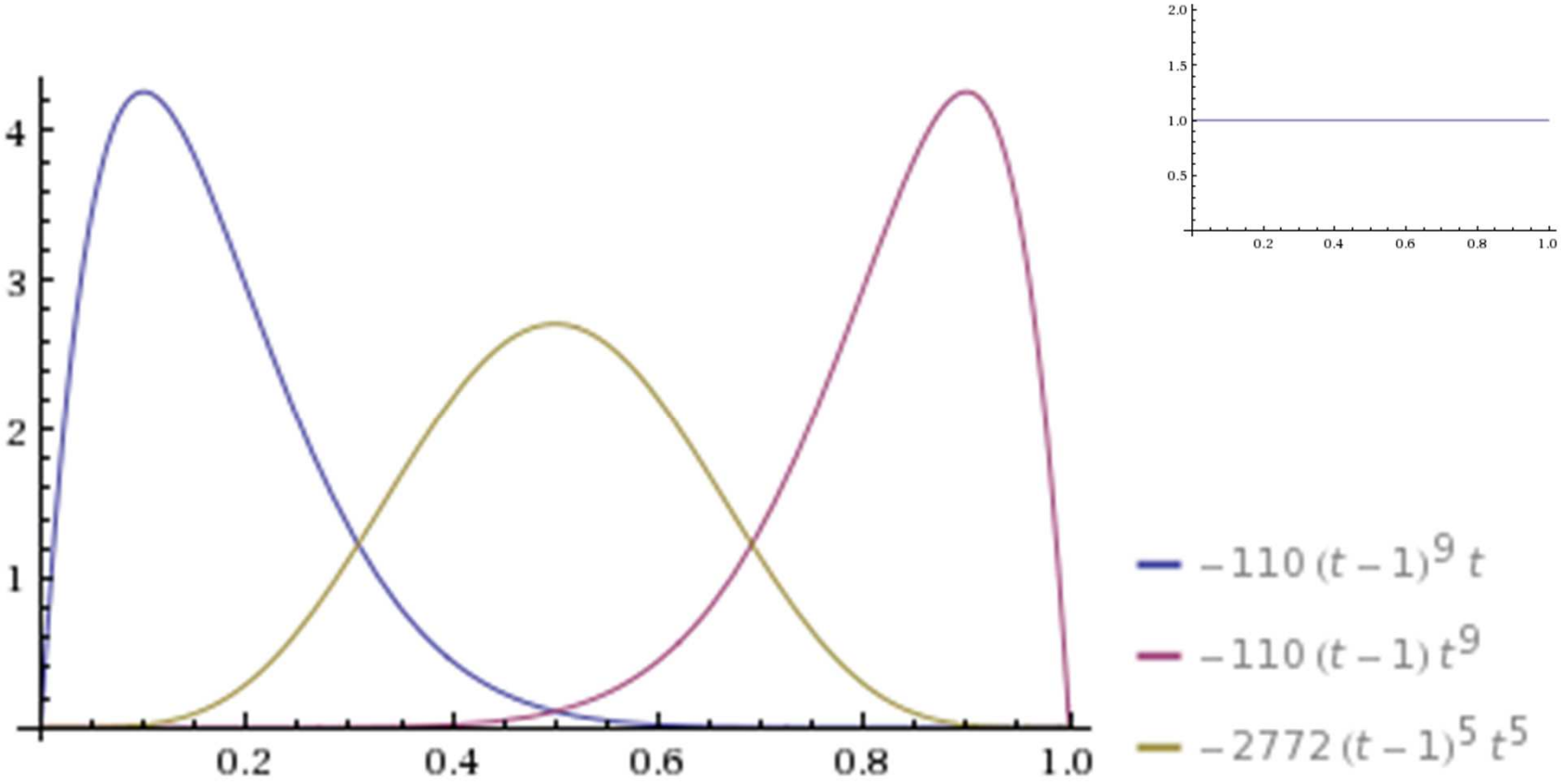
Beta(α,β) distr with mean = $\alpha/(\alpha+\beta)$

then the posterior distribution:

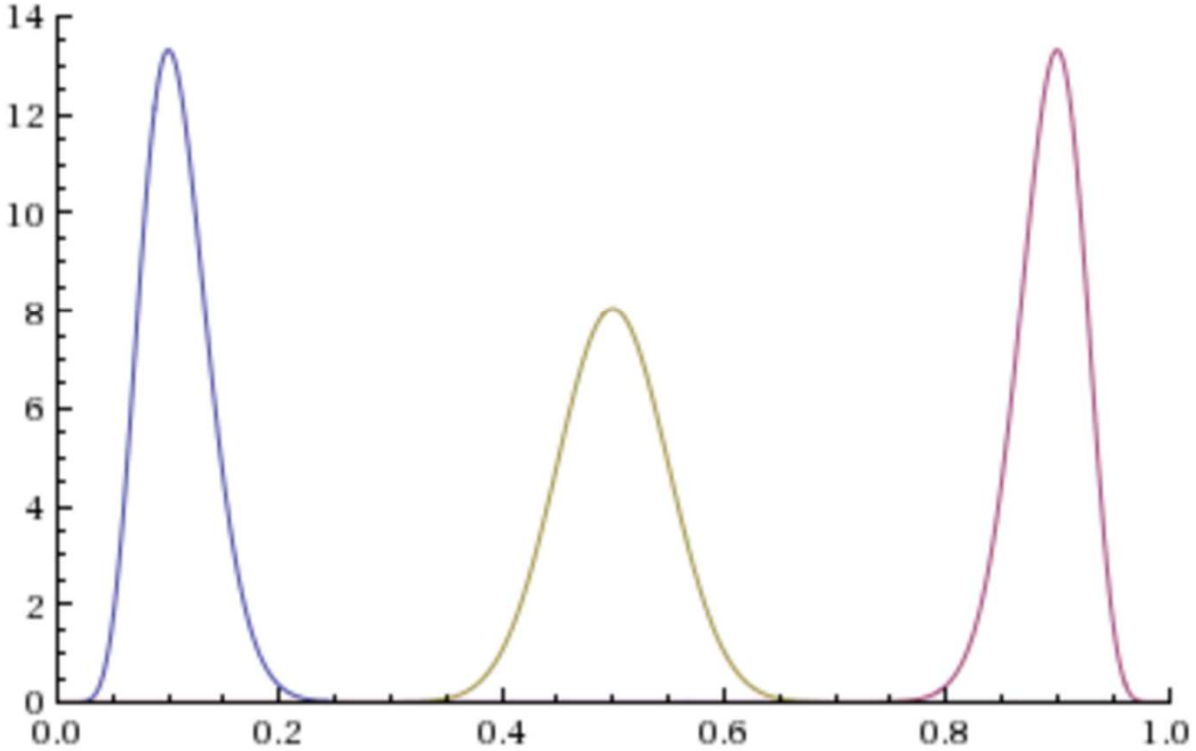
$$\text{Beta}\left(\sum_{i=1}^n x_i + \alpha, n - \sum_{i=1}^n x_i + \beta\right)$$

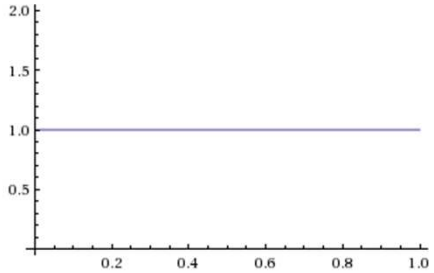
conjugate prior for Bernoulli distr.

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

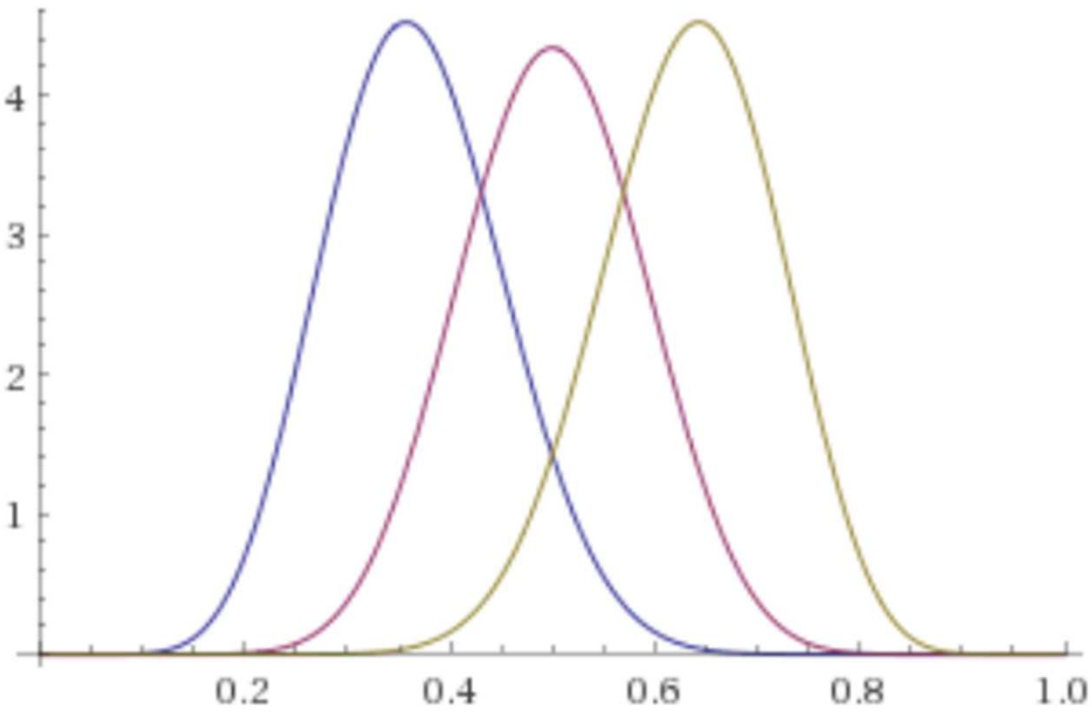


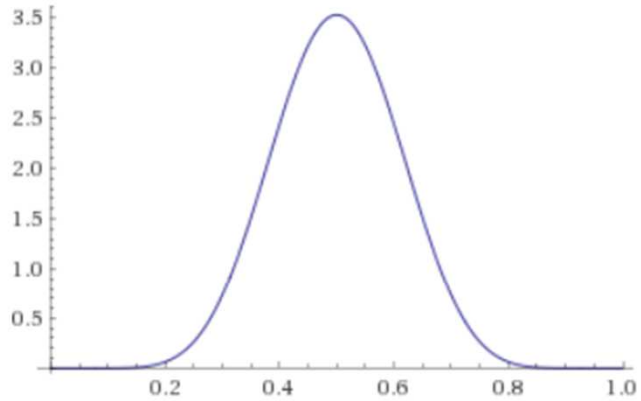
For a Beta (1,1) prior and data: n=100 and 10, 50, 90 successes





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



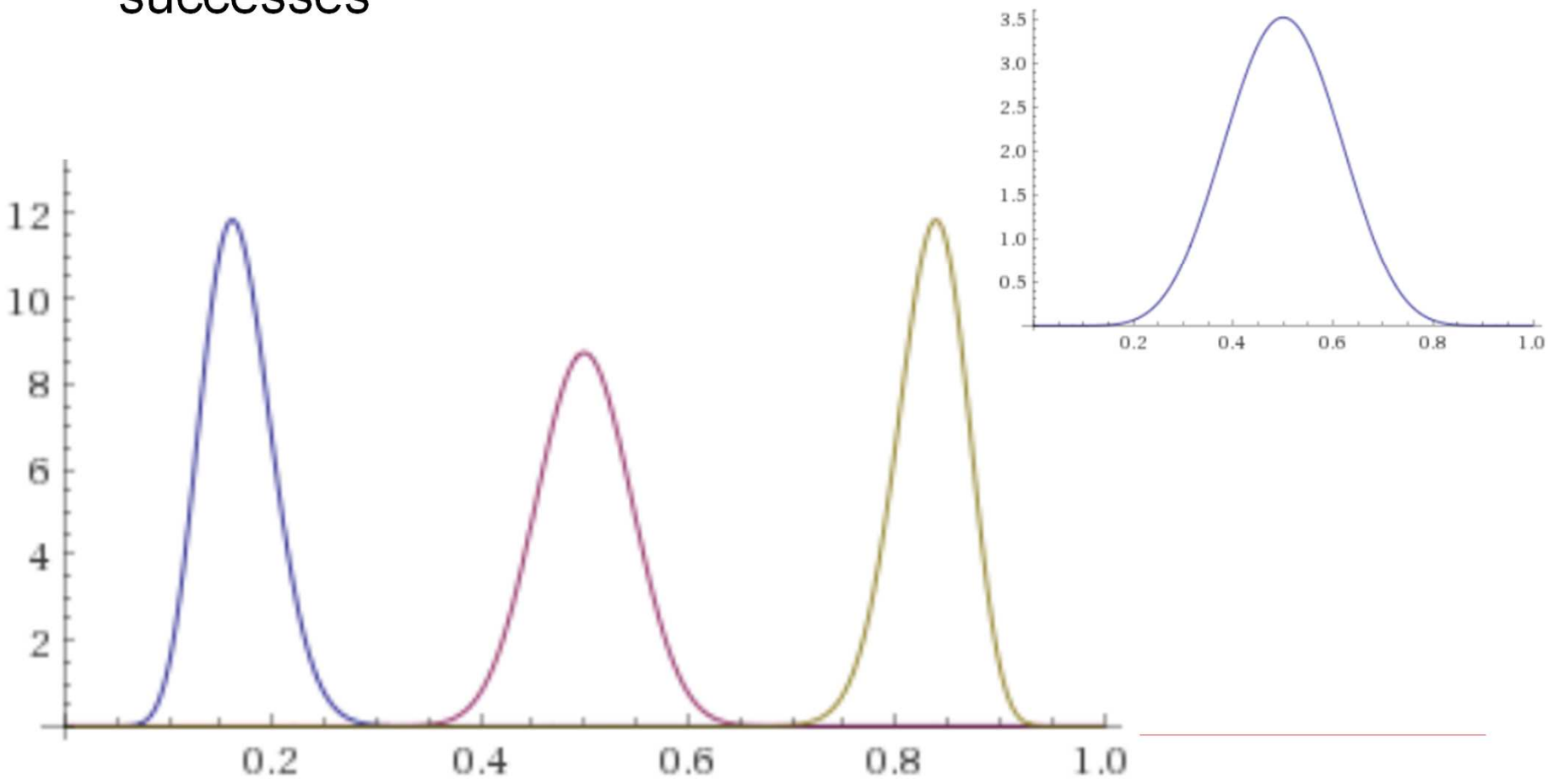


$380\,570\,190\,(x-1)^{18}\,x^{10}$

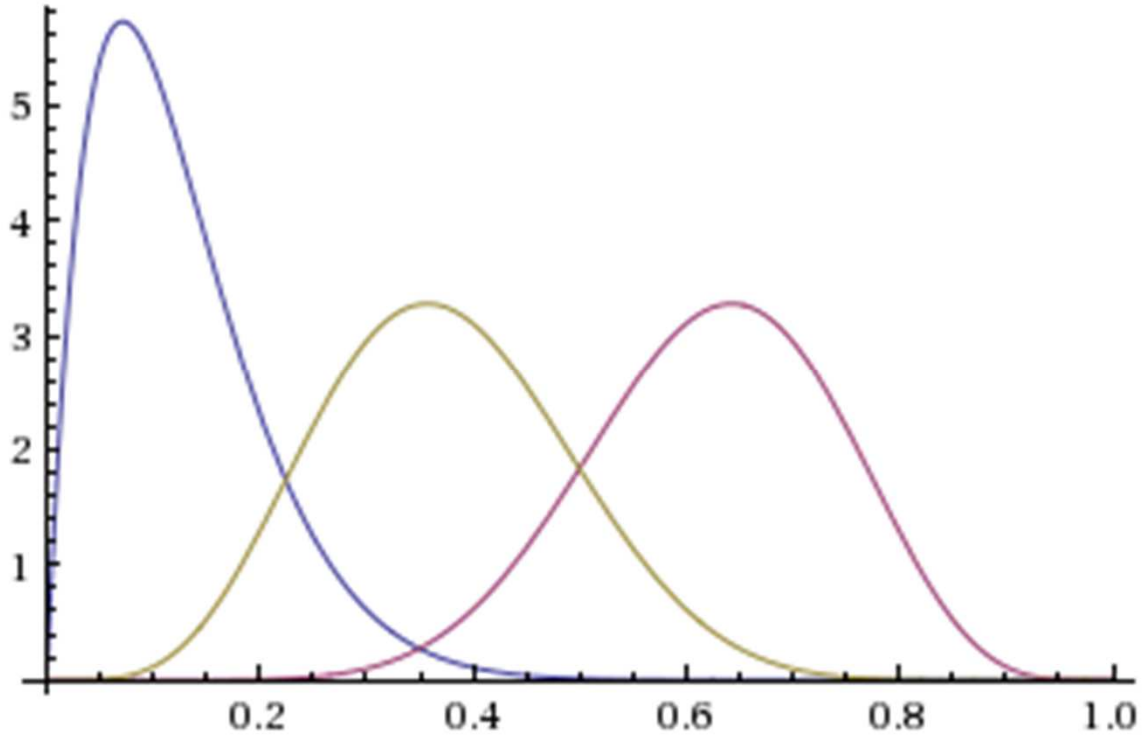
$1\,163\,381\,400\,(x-1)^{14}\,x^{14}$

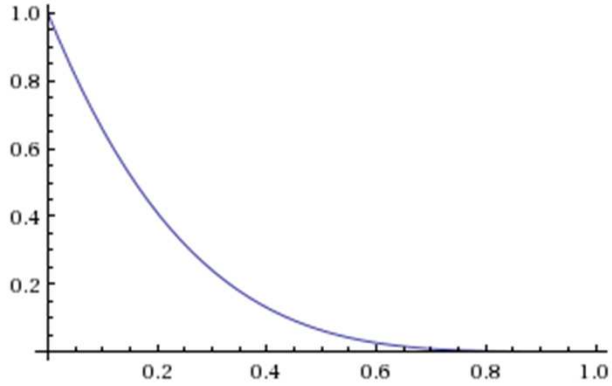
$380\,570\,190\,(x-1)^{10}\,x^{18}$

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



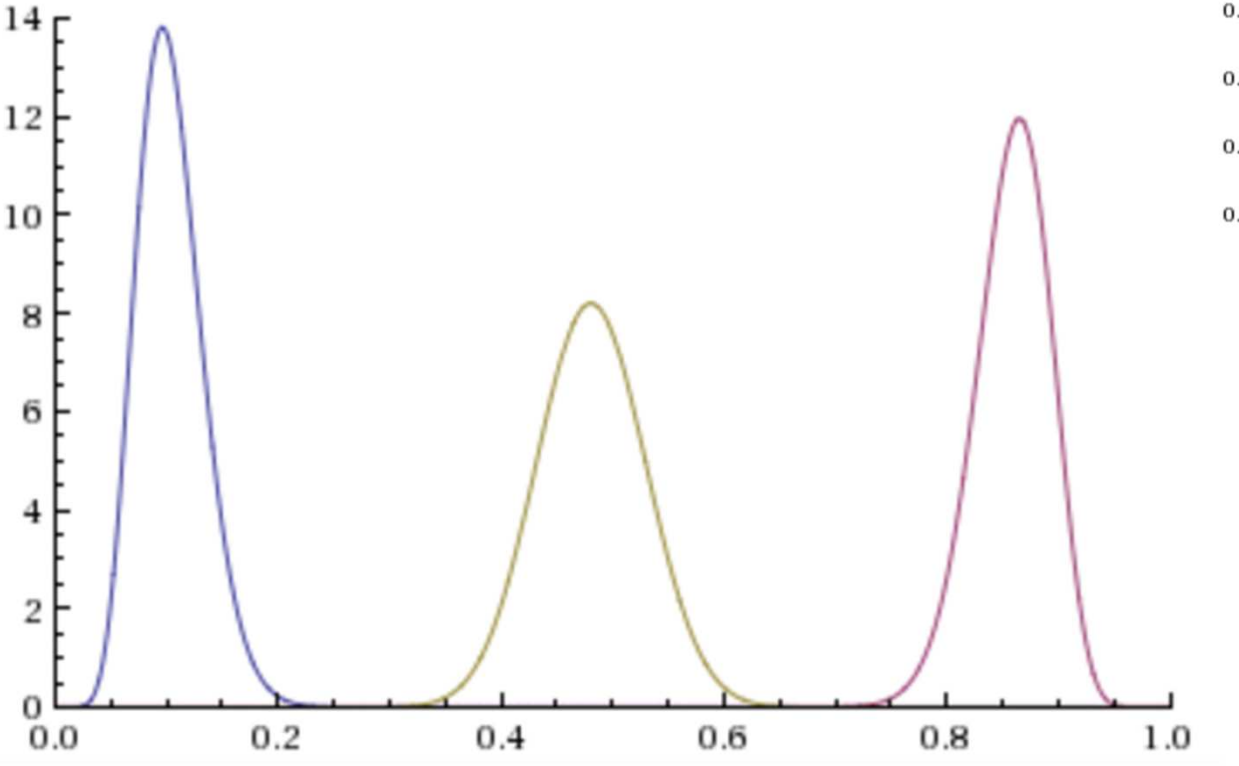


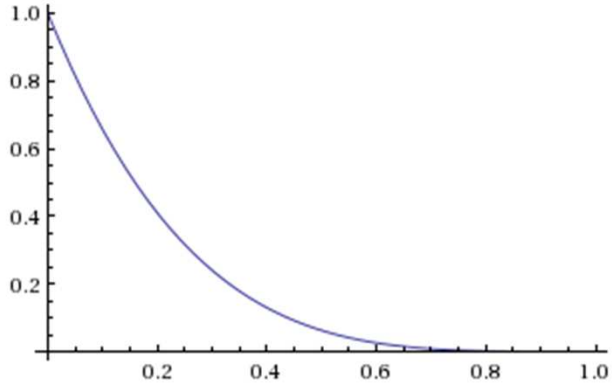
$-210\,(t-1)^{13}\,t$

$-30\,030\,(t-1)^{5}\,t^{9}$

$-30\,030\,(t-1)^{9}\,t^{5}$

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





A priori and a posteriori distributions: examples (2)

2. Let $X_1, \ldots, X_n$ 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:

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

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:

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

i.e.

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

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:

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

we have max for

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

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

i.e. for 5 successes in 10 trials for an a priori 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 a priori distr., we have *BMP*(θ)=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: $N\left(\frac{n\frac{1}{\sigma^2}\overline{X} + \frac{1}{\tau^2}m}{n\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{n\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)$

so

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

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 a priori 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 a priori 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 :

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

(the average loss of the estimator for a given a priori 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

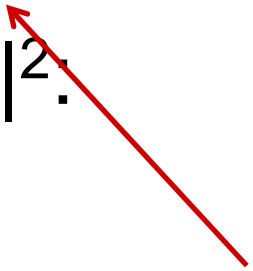
$$\forall x \quad acc(\Pi, \hat{g}_B(x)) = \min_a acc(\Pi, a)$$

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

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

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

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



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 $\theta \in (0,1)$

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

We know the posterior distribution:

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

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

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 an a priori 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 a priori 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 $\theta \in (0,1)$

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

We know the poster distribution:

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

we have max for

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

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

i.e. for 5 successes in 10 trials for an a priori 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 a priori 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: $N\left(\frac{n\frac{1}{\sigma^2}\overline{X} + \frac{1}{\tau^2}m}{n\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{n\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)$

so

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

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 a priori 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 a priori 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 $X_1, ..., X_n$ 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 : $N\left(\frac{n\frac{1}{\sigma^2}\overline{X} + \frac{1}{\tau^2}m}{n\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{n\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)$

so

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

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 a priori distr is $\theta \sim N(1, 1)$, then

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

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

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