Bayesian Statistics
Can we count on it ?

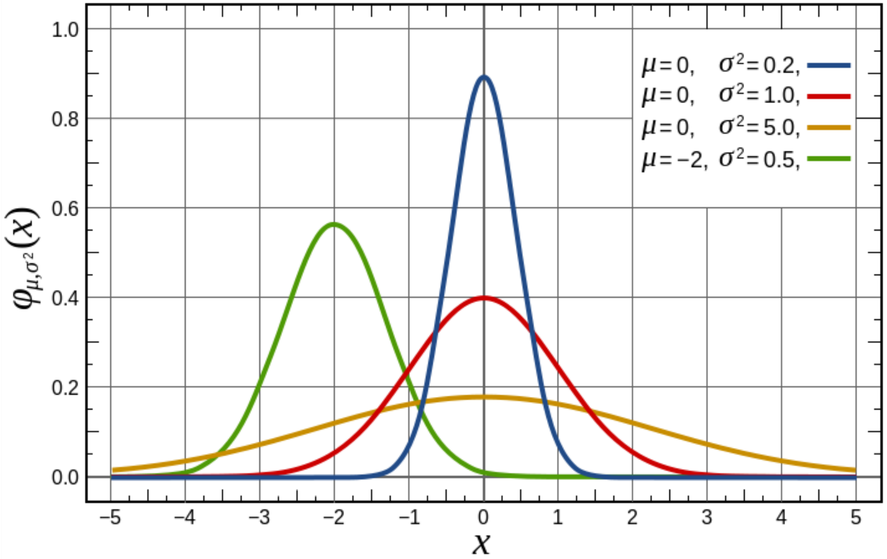
Uwe Menzel, 2012

uwe.menzel@matstat.de

www.matstat.org

Inference

- drawing conclusions from data with random variation (noise)
- more specific: infer parameters on the basis of samples



$\Downarrow$

μ, σ

Overview

- Basics, by means of 2 examples:
  - Table game (Thomas Bayes)
  - comparison with Maximum Likelihood
  - Coin flipping
  - comparison with Maximum Likelihood
- Empirical Bayes (EB)
  - edgeR and relatives

Related readings

PRIMER

What is Bayesian statistics?

Sean R Eddy

There seem to be a lot of computational biology papers with 'Bayesian' in their titles these days. What's distinctive about 'Bayesian' methods?

There are excellent introductory books on Bayesian analysis[1–3], but the key ideas behind the buzzword can be grasped quickly. Consider the following gambling puzzle—one

If *p* were known, this would be easy

Because Alice just needs one more point to win, Bob only wins the game if he takes the next three points in a row. The probability of

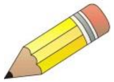
Inferring *p* from the data

The problem is that Alice and Bob don't know *p*. The very fact that Alice is ahead 5-3 is evidence that the unknown position of the mark

- decsription of the table game

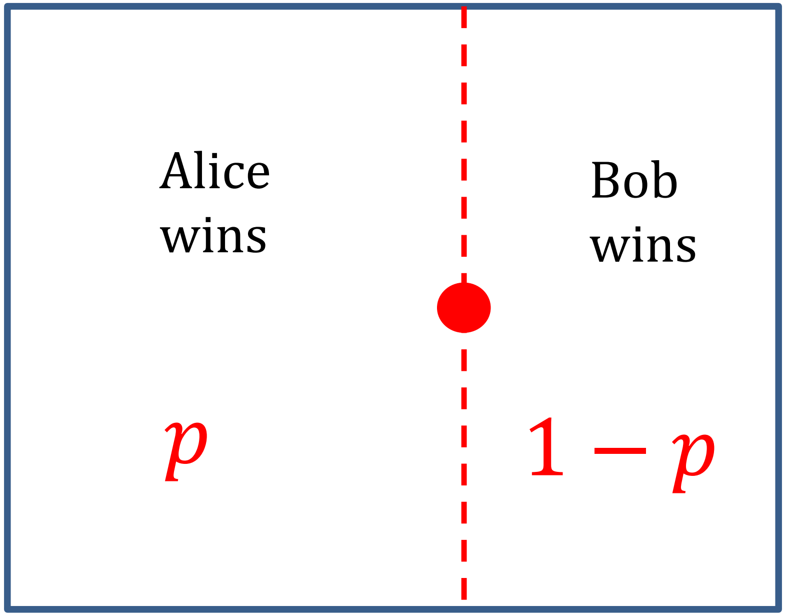
Essentials

- Binomial distribution, $Bin(n, p)$
- Expectation values, $E[X]$, $E[f(X)]$
- Bayes theorem (conditional probabilities)



Essentials.pdf

Table game: throw a ball



- Initial throw determines p – Alice and Bob don’t see it !
- Probability that Alice wins a single throw : p
- Probability that Bob wins a single throw : $1 - p$
- First player with 6 points wins
- **Intermediate result**: $A = 5; B = 3$
- How can Alice estimate her chances to win ?

Alice' odds

Intermediate result: $A = 5; B = 3$; first player with 6 points wins →
Bob can only win the game when he wins the next 3 throws :

$$P(Bob\ wins) = P(BBB) = (1 - p)^3$$

Alice wins if Bob does **not** win:

$$P(Alice\ wins) = 1 - P(Bob\ wins) = 1 - (1 - p)^3$$

(This is the easiest way to think of it since there are multiple possibilities how Alice can win.)

Hurray !!- that's it (the solution)!

... Is it ?
We don't have p !

1. The naive approach

Alice won 5 out of 8 throws → the probability that she wins in a single throw is 5/8:

$$A = 5\,; B = 3 \quad \Rightarrow \quad p = \frac{5}{8}$$

The probabilities to win the whole game are therefore:

$$P(Alice\ wins) = 1 - (1-p)^3 = 1 - \left(\frac{3}{8}\right)^3 = \frac{485}{512}$$

$$P(Bob\ wins) = \frac{27}{512}$$

$$odds = \frac{P(Alice\ wins)}{P(Bob\ wins)} \approx 18{:}\,1$$

2. Maximum-Likelihood (ML)

The game is a sequence of independent trials (Bernoulli trials); the probability of success in each trial is p. Therefore, the number of successes in n trials is binomially distributed:

$$P(k\ successes\ |\ p) = \binom{n}{k} p^k\,(1-p)^{n-k}$$

Probability mass function for the binomial distribution with probability of success = p

$$P(A=5; B=3\ |\ p) = \binom{8}{5} p^5\,(1-p)^3$$

Probability that Alice wins 5 throws out of 8, probability p to win a single throw unknown

In ML, we search for the parameter p that makes the observation most likely, i.e. we maximize the following expression w.r.t. the parameter p:

$$L(p) = \binom{8}{5} p^5\,(1-p)^3 \Longrightarrow \text{Maximum}$$

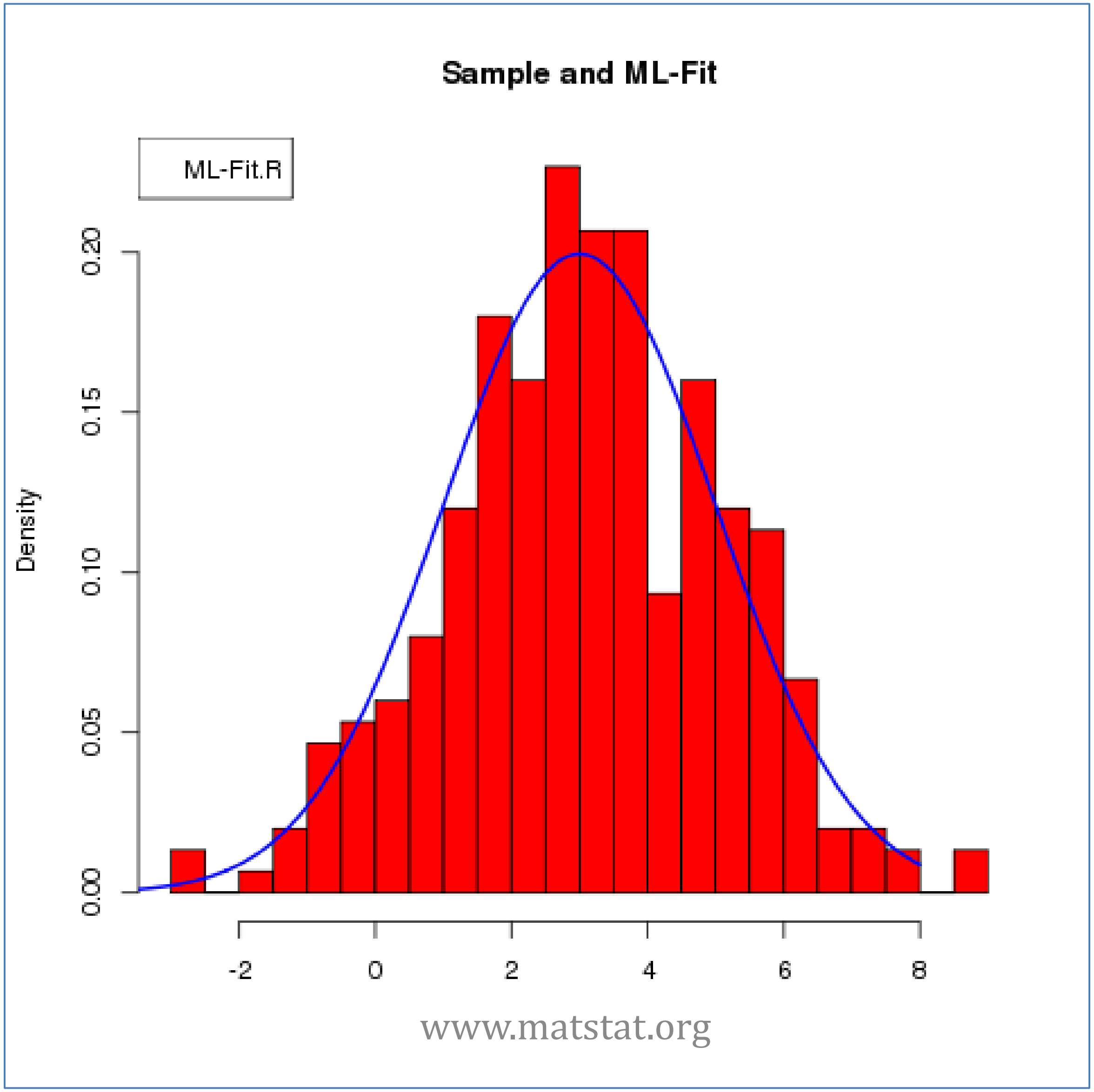
$$l(p) = ln(L) = C + 5 \cdot ln(p) + 3 \cdot ln(1-p)$$

we can maximize the logarithm instead (easier)

$$\frac{dl}{dp} = \frac{5}{p} - \frac{3}{1-p} = 0 \Longrightarrow p = \frac{5}{8} \Longrightarrow odds \approx 18{:}1$$

(as in the naive approach)

2. Maximum-Likelihood (ML)



3. Bayesian approach

We had: $P(Bob\ wins) = P(BBB) = (1-p)^3$. Now, the idea is to calculate the expected value of this expression by considering p as a random variable:

$$E(Bob\ wins) = E\left[(1-p)^3\right]$$ expectation, p random!

Because p is continuous in the interval $(0, 1)$, this reads:

$$E\left[(1-p)^3\right] = \int_0^1 (1-p)^3 \cdot f(p)\, dp$$

Here, a probability density function $f(p)$ was introduced. This stands for the main idea of the Bayesian approach: we treat the parameter under investigation as a random variable, i.e. we allow the parameter p to be distributed with some $f(p)$. The observation made is incorporated into the calculation by using for $f(p)$ the conditional probability, conditioned on the observed data, $f(p) = P(p \mid A = 5, B = 3)$, so that we get:

$$E\left(Bob\ wins\right) = \int_0^1 (1-p)^3 \cdot P(p \mid \underbrace{A = 5, B = 3}_{\text{observed data}})\, dp$$

3. Bayesian approach

$$E\,(Bob\ wins) = \int_0^1 (1-p)^3 \cdot P(p \mid A=5, B=3)\; dp$$

We need $P(p\,|A=5; B=3)$, the probability distribution of the parameter p given the observed data. This is called the posterior probability, because it is a probability determined **after** seeing the data. However, we don't have $P(p \mid A=5; B=3)$, we have only $P(A=5; B=3 \mid p)$, delivered by the binomial probability mass function. This is a nice chance to use Bayes law:

$$P\,(p \mid 5,3) = \frac{P(5,3 \mid p) \cdot P(p)}{P(5,3)}$$

Here, $P(p)$ is the unconditioned (prior) probability distribution of p, and $P(5,3) = P(A=5; B=3)$ is the total probability of the observation. The latter can be calculated using the Law of total probability, leading to:

$$P\,(p \mid 5,3) = \frac{P(5,3 \mid p) \cdot P(p)}{\int_0^1 P(5,3 \mid p) \cdot P(p)}$$

3. Bayesian approach

$$E\,(Bob\ wins) = \int_0^1 (1-p)^3 \cdot P(p \mid 5,3)\ dp$$ now use Bayes law →

$$E\,(Bob\ wins) = \int_0^1 (1-p)^3 \cdot \frac{P(5,3 \mid p) \cdot P(p)}{P(5,3)}\ dp$$ now use total prob. →

$$E\,(Bob\ wins) = \frac{\int (1-p)^3 \cdot P(5,3 \mid p) \cdot P(p)\ dp}{\int P(5,3 \mid p) \cdot P(p)\ dp}$$

$P(p) = 1$
flat prior

We need the prior distribution $P(p)$. If we have no idea about this distribution, we might use a "flat prior", $P(p) = 1$ in $(0, 1)$, so that we get:

$$E\,(Bob\ wins) = \frac{\int (1-p)^3 \cdot \binom{8}{5}\ p^5\,(1-p)^3\ \ dp}{\int \binom{8}{5}\ p^5\,(1-p)^3\ \ dp}$$ PMF of the binomial distribution was used here

$$E\,(Bob\ wins) = \frac{\int_0^1 p^5\,(1-p)^6\ \ dp}{\int_0^1 p^5\,(1-p)^3\ \ dp}$$ integral can be solved

3. Bayesian approach

$$E\,(Bob\ wins) = \frac{\int_0^1 p^5\,(1-p)^6\ dp}{\int_0^1 p^5\,(1-p)^3\ dp}$$

Beta integral, leads to Gamma function →

$$\int_0^1 p^{m-1}\cdot(1-p)^{n-1}\ dp = \frac{\Gamma(n)\cdot\Gamma(m)}{\Gamma(n+m)} = \frac{(n-1)!\cdot(m-1)!}{(n+m-1)!}$$

$$\Rightarrow\quad E\,(Bob\ wins) = \frac{\int_0^1 p^5\,(1-p)^6\ dp}{\int_0^1 p^5\,(1-p)^3\ dp} = \frac{5!\cdot 6!\cdot 9!}{12!\cdot 5!\cdot 3!} = \frac{1}{11}$$

$$\Rightarrow\quad E(Alice\ wins) = \frac{10}{11}$$

$$\Rightarrow\quad odds(Alice\ wins) = 10:1$$

Comparison of the results

$$odds = \frac{P(Alice\ wins)}{P(Bob\ wins)} = 18:1$$ naïve approach

$$odds = \frac{P(Alice\ wins)}{P(Bob\ wins)} = 18:1$$ Maximum Likelihood

$$odds = \frac{P(Alice\ wins)}{P(Bob\ wins)} = 10:1$$ Bayesian approach

Which one is correct ?

Which one is correct?

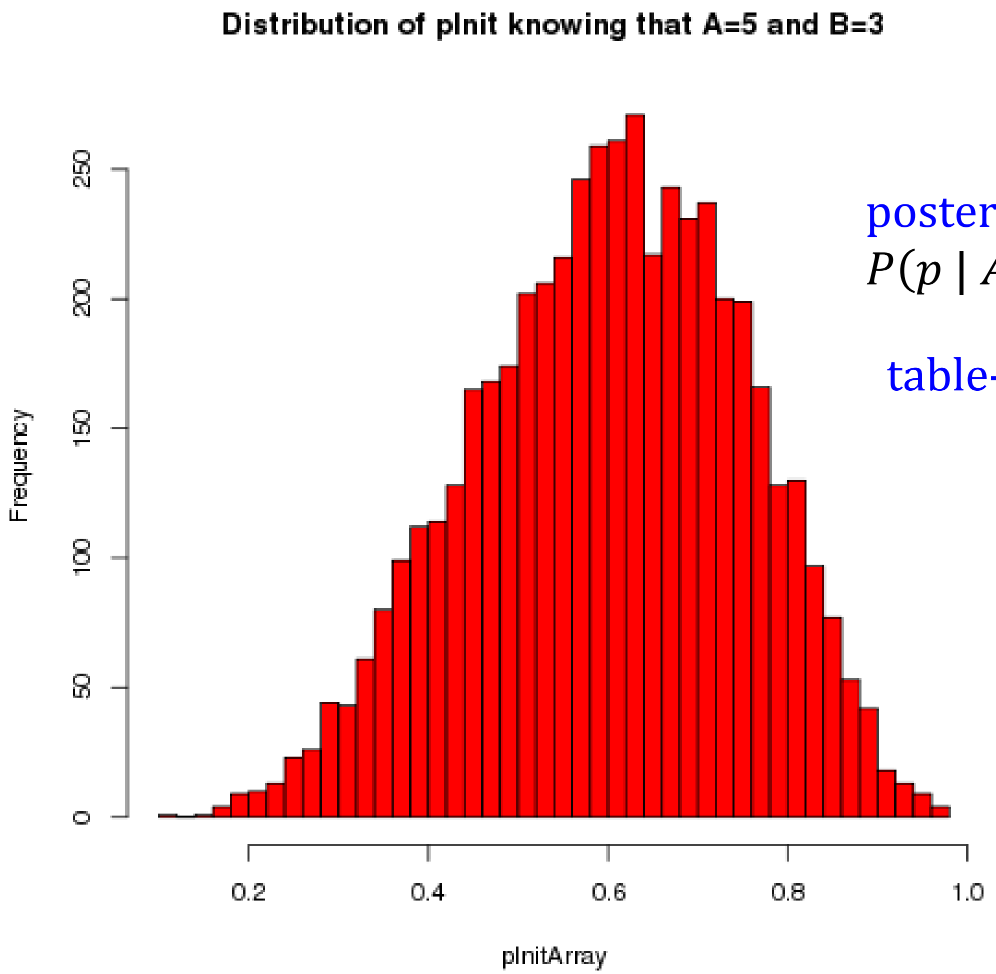
- Just play the game (a lot of times)
- see table_Game.html; more details in table_Game.R

```
NumberAliceWins = 0
NumberBobWins = 0
numberGames = 5000
pInitArray = numeric(numberGames)

for (i in 1:numberGames) {
  pInit = get_pInit()          # renew in each game!
  pInitArray[i] = pInit        # save for histogram of posterior distribution
  AlicePoints = 5              # current score
  BobsPoints = 3

  while ( (AlicePoints < 6) && (BobsPoints < 6)) {  # play this game until one participant wins
    nextThrow = runif(1, min = 0, max = 1)
    if ( nextThrow <= pInit) {AlicePoints = AlicePoints + 1} else {BobsPoints = BobsPoints + 1}
  }
  if(AlicePoints == 6) {NumberAliceWins = NumberAliceWins + 1} else {NumberBobWins = NumberBobWins + 1}
}
(NumberAliceWins + NumberBobWins) == numberGames  # This must be TRUE
```

Uwe Menzel, 2012

Distribution of the posterior probability

- given the intermediate result A=5 & B=3 -

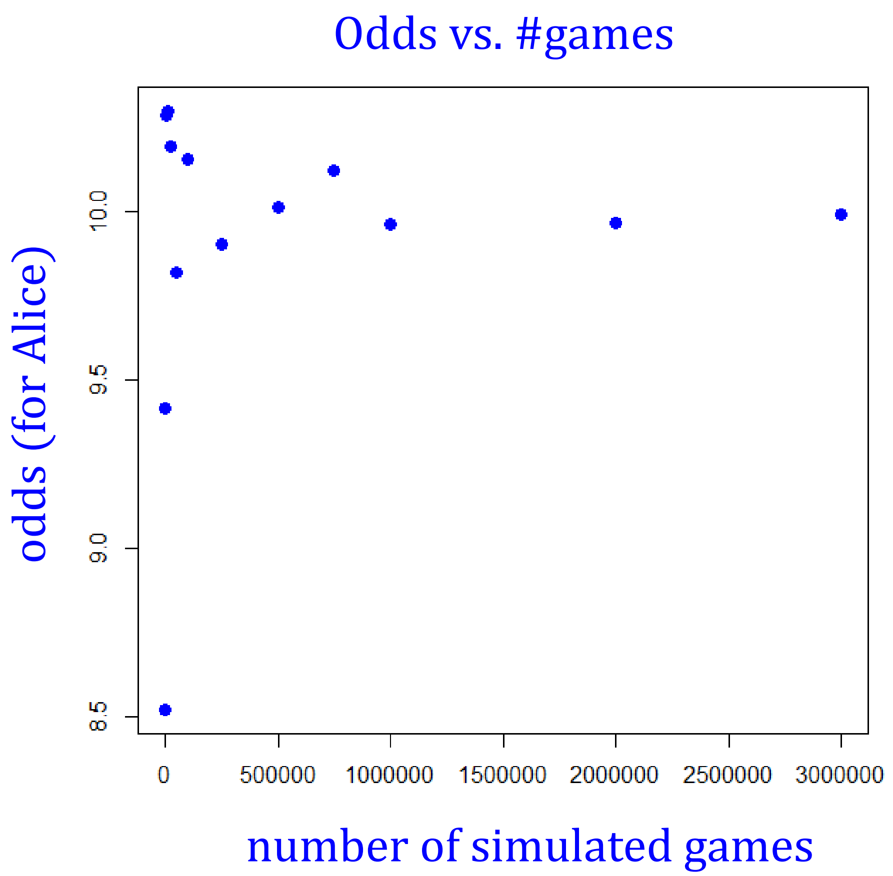


posterior probability
$P(p \mid A = 5, B = 3)$

table-Game.R

Results of the table game simulation

- see table-Game.html (linked)
- the table game algorithm includes random components
- → better and better results can be achieved by simulating more and more games
- → increase number of played games until the results get stable:



The simulation yields the odds 10: 1, the result of the Bayesian approach.

Coin flipping

head, probability $= p$

tail, $1 - p$

- **Task**: infer p (which might not be exactly 0.5 !)
- Use:
  - observed data: number of heads tossed; number of tails tossed
  - a-priori knowledge (experience): p should be close to 0.5

Naive approach:

- 10 flips → $h = 3; t = 7$ (ten flips are by far not enough, but let us use this for now to demonstrate the principle)
- → $P(h) = 3/10\ ; P(t) = 7/10$

Hmmh, I don't think we can trust that, this is too far from 0.5. It contradicts experience. Try Maximum Likelihood →

Maximum Likelihood

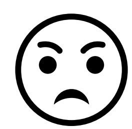
Let p be the probability to flip head ("success"). A single flip can be regarded as a Bernoulli trial. The number of successes in n independent Bernoulli trials is binomially distributed, with the probability mass function

$$P\left(3 \text{ heads in } 10 \text{ casts}\right) = \binom{10}{3} p^3 \left(1-p\right)^7$$

$$L(p) = \binom{10}{3} p^3 \left(1-p\right)^7$$ Likelihood, to maximize!

$$l(p) = \ln(L) = C + 3 \cdot \ln(p) + 7 \cdot \ln(1-p)$$

$$\frac{dl}{dp} = \frac{3}{p} - \frac{7}{1-p} = 0 \quad \Longrightarrow \quad p = P\left(\text{heads}\right) = \frac{3}{10}$$



Bayesian approach

We search the probability p to flip head. As in the previous example, we calculate the expected value of this parameter by treating p as a random variable:

$$E(p) = \int_0^1 p \cdot P(p \mid data)\ dp$$

expectation, p random

Again, we use a distribution of p which is conditioned on the observed data. Using Bayes law, we can write:

posterior probability → likelihood → prior probability

$$P(p \mid data) = \frac{P(data \mid p) \cdot P(p)}{P(data)}$$

alternative: MAP (maximum a posteori)

posterior ~ likelihood x prior

Bayesian approach

posterior probability → likelihood → prior probability

$$P(p \mid data) = \frac{P(data \mid p) \cdot P(p)}{P(data)}$$

posterior ~ likelihood x prior

$$P(data \mid p) = \binom{10}{3} p^3 (1-p)^7$$ likelihood, from binomial distribution

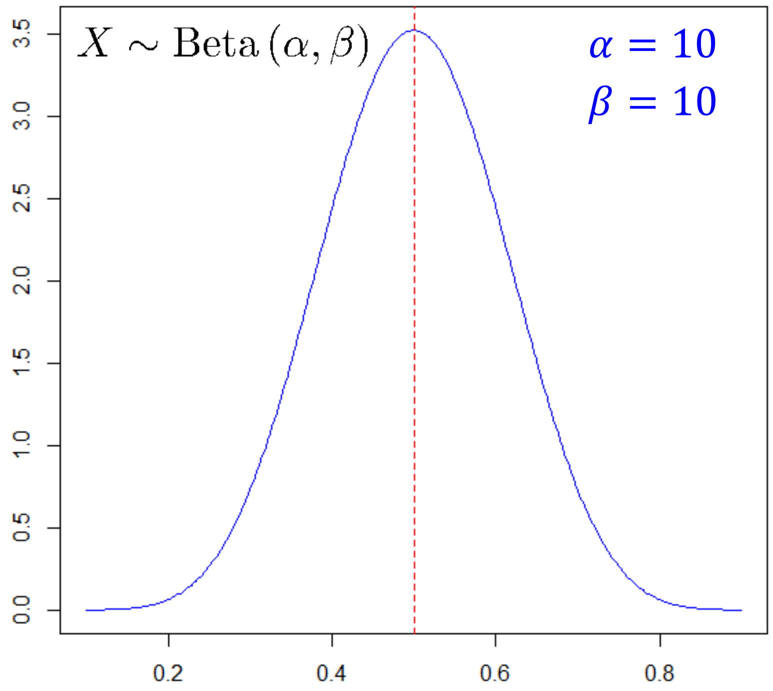
$P(p)$: prior distribution, to be chosen. See below.

$P(p \mid data)$: posterior distribution, can be calculated once prior is chosen. See below.

The prior distribution $P(p)$

- **Bayesian inference**: consider $p = P(head)$ as not being "sharp", but distributed with some probability density function
- based on experience, we expect p to be closely distributed around 0.5
- therefore, we choose a prior that is concentrated around 0.5
- using the Beta-distribution is very convenient, as we will see below

Probability Density Function



function `dbeta()` in R

$$X \sim Beta(\alpha, \beta)$$

$$E(X) = \frac{\alpha}{\alpha + \beta}$$

$$V(X) = \frac{\alpha \cdot \beta}{(\alpha + \beta)^2 \cdot (\alpha + \beta + 1)}$$

The prior distribution $P(p)$

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

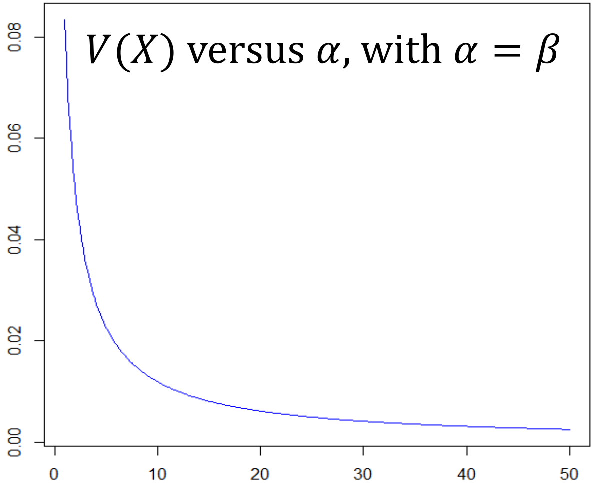
$$E(X) = \frac{\alpha}{\alpha+\beta} \qquad V(X) = \frac{\alpha\cdot\beta}{(\alpha+\beta)^2\cdot(\alpha+\beta+1)}$$ mean and variance

$$\alpha = \beta \;\rightarrow\; E(X) = 0.5$$

mean is 0.5, as desired for coin flipping

$$\alpha = \beta \;\rightarrow\; V(X) \sim \frac{1}{8\cdot\alpha+4}$$

The bigger α and β (with $\alpha = \beta$), the lower the variance → possibility to control the shape of the prior. $\alpha = \beta = 100 \;\rightarrow V(X) = 0.0012$



α and β are called hyperparameters, because they determine the distribution of another parameter: p

The Beta-distribution (with $\alpha = \beta$) seems to be a suitable prior for the coin-flipping problem

The posterior distribution $P(p \mid data)$

$$P(p \mid data) = \frac{P(data \mid p) \cdot P(p)}{P(data)}$$

posterior ; likelihood ; prior

$$P(p \mid data) = \frac{1}{P(data)} \cdot \binom{10}{3} p^3 \, (1-p)^7 \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$

$$P(data) = \int P\,(data \mid p) \cdot P(p)\; dp$$

law of total probability

$$= \int_0^1 \binom{10}{3} p^3 \, (1-p)^7 \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1} \;\; dp$$

$$\Rightarrow \quad P\,(p \mid data) = \frac{p^{3+\alpha-1} \cdot (1-p)^{7+\beta-1}}{\int_0^1 p^{3+\alpha-1} \cdot (1-p)^{7+\beta-1} \;\; dp}$$

in general, we have $\displaystyle \int_0^1 p^{m-1} \cdot (1-p)^{n-1} \;\; dp = \frac{\Gamma(m) \cdot \Gamma(n)}{\Gamma(m+n)}$

The posterior distribution $P(p \mid data)$

in general, we have $$\int_0^1 p^{m-1} \cdot (1-p)^{n-1} \; dp = \frac{\Gamma(m) \cdot \Gamma(n)}{\Gamma(m+n)}$$

which yields $$\int_0^1 p^{3+\alpha-1} \cdot (1-p)^{7+\beta-1} \; dp = \frac{\Gamma(3+\alpha) \cdot \Gamma(7+\beta)}{\Gamma(10+\alpha+\beta)}$$

$$P\left(p \mid data\right) = \frac{\Gamma(10+\alpha+\beta)}{\Gamma(3+\alpha) \cdot \Gamma(7+\beta)} \cdot p^{3+\alpha-1} \cdot (1-p)^{7+\beta-1}$$

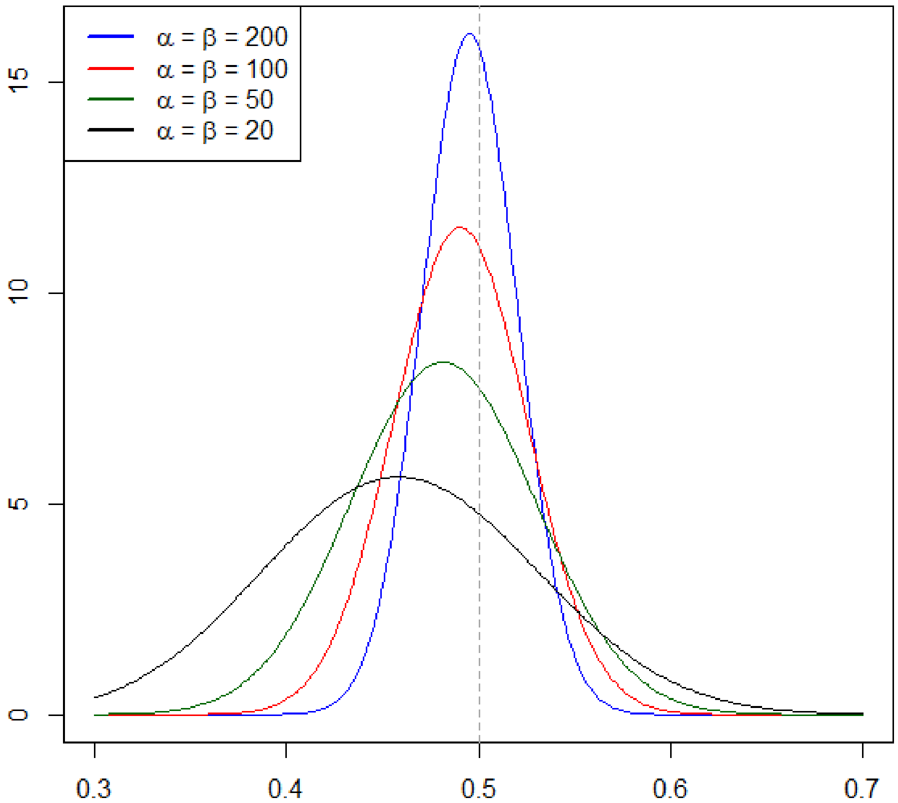
$$P\left(p \mid data\right) \sim Beta\left(3+\alpha, 7+\beta\right)$$ posterior is Beta-distributed

because $$Beta\left(p \mid m, n\right) = \frac{\Gamma(m+n)}{\Gamma(m) \cdot \Gamma(n)} \cdot p^{m-1} \cdot (1-p)^{n-1}$$ PDF $m = 3 + \alpha$, $n = 7 + \beta$

We used as prior the distribution $Beta(\alpha, \beta)$. The posterior distribution is also a Beta distribution with somewhat changed parameters. As we have seen above, we can arbitrarily narrow down the posterior by choosing higher and higher values for α and β (see also next page). Furthermore, as we will see soon, this also shifts the expectation for p towards the value 0.5.

The posterior distribution $P(p \mid data)$

We can arbitrarily narrow down the posterior by choosing higher and higher values for α and β. That also shifts the expectation for p towards the value 0.5 (see below).



PDF of the calculated posterior probability $Beta(3+\alpha, 7+\beta)$ for different values of α and β, with $\alpha = \beta$. Higher values of α and β confine the posterior to the region around the mean, which is 0.5 (if $\alpha = \beta$).

Result: Expectation of p

$$E(p) = \int_0^1 p \cdot \underbrace{P\left(p \mid data\right)}_{\text{Posterior distribution}} \; dp$$ expectation, p random

$$P\left(p \mid data\right) = \frac{\Gamma(10+\alpha+\beta)}{\Gamma(3+\alpha)\cdot\Gamma(7+\beta)} \cdot p^{3+\alpha-1} \cdot (1-p)^{7+\beta-1}$$ posterior distribution

$$E(p) = \frac{\Gamma(10+\alpha+\beta)}{\Gamma(3+\alpha)\cdot\Gamma(7+\beta)} \cdot \int_0^1 p^{4+\alpha-1} \cdot (1-p)^{7+\beta-1} \; dp$$

use: $$\int_0^1 p^{m-1} \cdot (1-p)^{n-1} \; dp = \frac{\Gamma(m)\cdot\Gamma(n)}{\Gamma(m+n)}$$

$$E(p) = \frac{\Gamma(10+\alpha+\beta)}{\Gamma(3+\alpha)\cdot\Gamma(7+\beta)} \cdot \frac{\Gamma(4+\alpha)\cdot\Gamma(7+\beta)}{\Gamma(11+\alpha+\beta)}$$

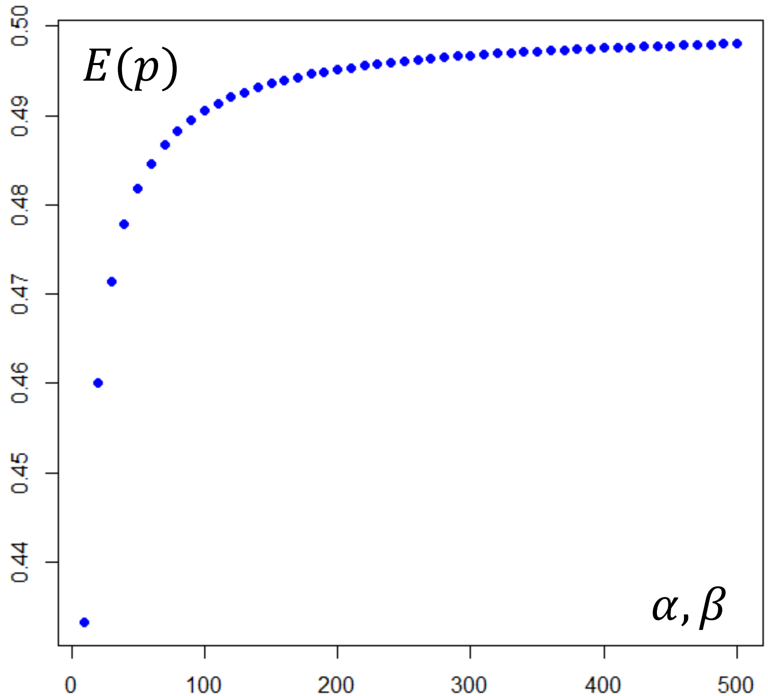
$$E(p) = \frac{\Gamma(10+\alpha+\beta)}{\Gamma(3+\alpha)} \cdot \frac{\Gamma(4+\alpha)}{\Gamma(11+\alpha+\beta)}$$

Result: Expectation of p

$$E(p) = \frac{\Gamma(10+\alpha+\beta)}{\Gamma(3+\alpha)} \cdot \frac{\Gamma(4+\alpha)}{\Gamma(11+\alpha+\beta)}$$

This is easier to calculate if we choose integers for α and β. In this case, we can use $\Gamma(n) = (n-1)!$ (the Γ-function for big arguments might be hard to calculate)

$$\Rightarrow \quad E(p) = \frac{(9+\alpha+\beta)!}{(2+\alpha)!} \cdot \frac{(3+\alpha)!}{(10+\alpha+\beta)!} = \frac{3+\alpha}{10+\alpha+\beta}$$



Increasing the hyperparameters α and β drives the solution of the coin flipping problem, i.e. the expected value of p, towards 0.5. By choosing appropiate values for α and β , we can come as close as desired to 0.5. This makes the Bayesian approach somewhat arbitrary! We can only choose hyperparameters which are well-established!

Summary, coin flipping experiment

Approach	Estimated p
Naïve approach	0.3
Maximum likelihood	0.3
Bayes, $\alpha = \beta = 100$	0.4905

- Applying the Bayesian approach, we have choosen a prior with a very narrow distribution around 0.5 ($\alpha = \beta = 100$).
- By incorporating the prior distribution, we actually add pseudocounts to the observed counts of both head and tail, driving the expectation for p towards 0.5.
- Adding more and more pseudocounts (higher α and β) assigns more and more weight to prior knowledge.
- We have to find a trade-off between the actually observed data and the prior knowledge (represented by the prior distribution).

RNA-Seq, Microarrays

Task: compare groups (healthy ↔ sick, treated ↔ untreated, …)

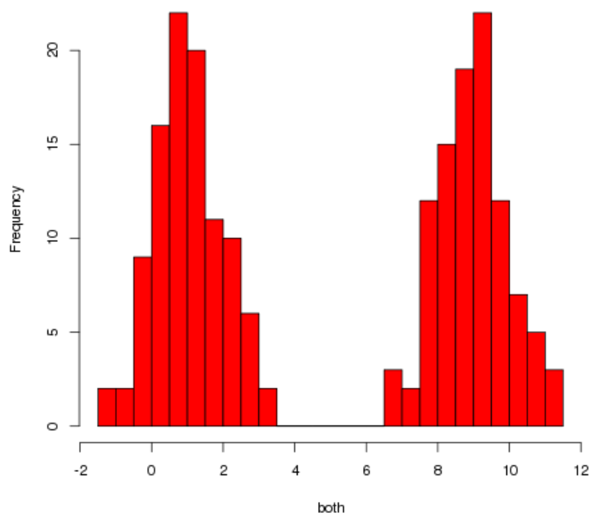
- find Differentially Expressed Genes (DEG's)
- Statistical (parametric) tests →

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}} \sim t(f)$$

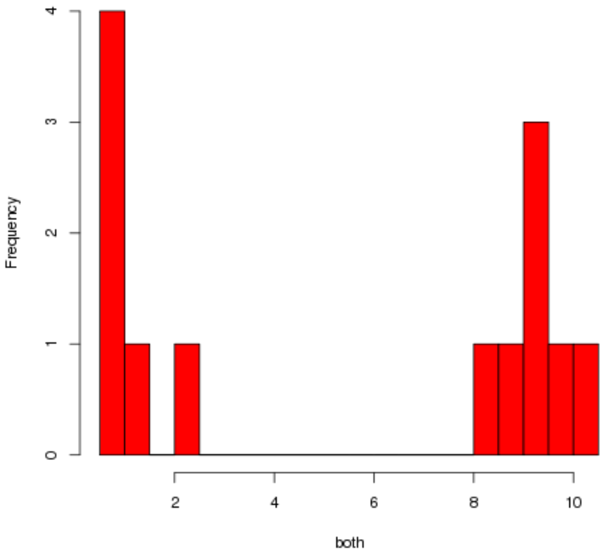
Problem: too few measurements in the groups

- unreliable estimates for the parameters $(\bar{X}, \bar{Y}, S_x^2, S_y^2)$
- hinders identification of significantly DEG's

large samples

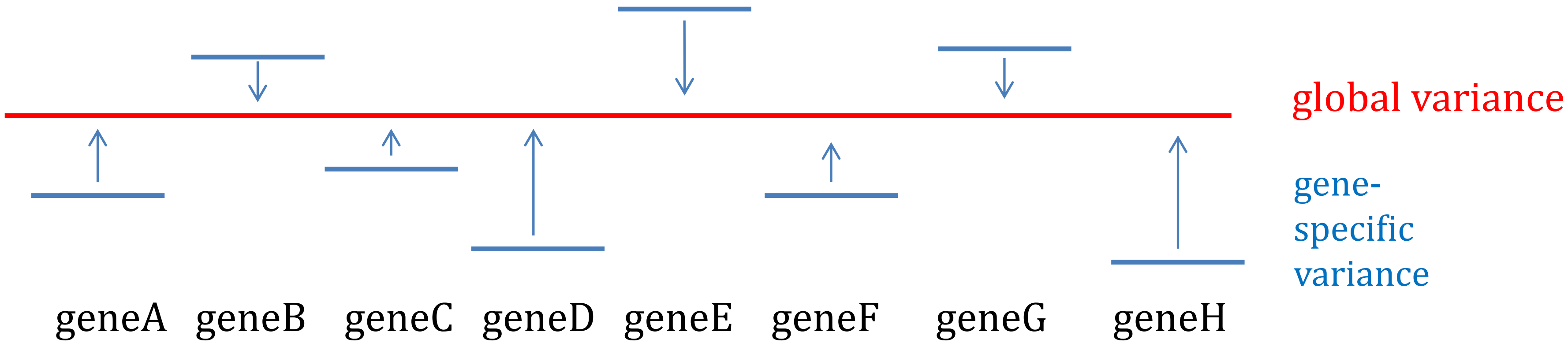


small samples



edgeR

- Robinson and Smyth, Biostatistics 2008
  - estimating the NegBin-variance (dispersion) globally across all genes
  - common dispersion across all genes
- Robinson and Smyth, Bioinformatics 2007
  - empirical Bayes model for variance estimation
  - permits gene specific dispersion which is though driven towards a common value inferred from all genes
- Robinson, McCarthy, Smyth, Bioinformatics 2010
  - edgeR , see the edgeR users guide



Standard Bayesian

$$E(p) = \int_0^1 p \cdot \frac{P(data\,|\,p) \cdot P(p)}{P(data)}\, dp$$

$P(p)$: Beta-function, parameters α and β

prior is choosen without looking at our own data (above, we have choosen α and β out of prior knowledge, not connected to the actual data observed)

Empirical Bayes (EB)

$$E(\varphi) = \int_0^1 \varphi \cdot \frac{P(data\,|\varphi) \cdot P(\varphi)}{P(data)}\, d\varphi$$

$P(\varphi)$: a function of parameters (which can in turn be parametrized)

hyperparameters estimated from the actual observation (e.g. borrowing information from neighboring locations in same dataset)

Appendix

Bayesian Statistics

Uwe Menzel, 2012

www.matstat.org

Conjugate priors for discrete random variables

Wikipedia

Discrete likelihood distributions [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	*p* (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^n x_i,\ \beta + n - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
Binomial	*p* (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^n x_i,\ \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x}\vert\alpha', \beta')$ (beta-binomial)
Negative Binomial with known failure number *r*	*p* (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^n x_i,\ \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures[note 1] (i.e. $\frac{\beta - 1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	$k,\ \theta$	$k + \sum_{i=1}^n x_i,\ \frac{\theta}{n\theta + 1}$	k total occurrences in $1/\theta$ intervals	$\operatorname{NB}(\tilde{x}\vert k', \frac{\theta'}{1+\theta'})$ (negative binomial)
Poisson	λ (rate)	Gamma	$\alpha,\ \beta$ [note 3]	$\alpha + \sum_{i=1}^n x_i,\ \beta + n$	α total occurrences in β intervals	$\operatorname{NB}(\tilde{x}\vert\alpha', \frac{1}{1+\beta'})$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), *k* (number of categories, i.e. size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category *i*	$\alpha_i - 1$ occurrences of category i[note 1]	$p(\tilde{x} = i) = \frac{\alpha_i{}'}{\sum_i \alpha_i{}'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$
Multinomial	$\boldsymbol{p}$ (probability vector), *k* (number of categories, i.e. size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + \sum_{i=1}^n \mathbf{x}_i$	$\alpha_i - 1$ occurrences of category i[note 1]	$\operatorname{DirMult}(\tilde{\mathbf{x}}\vert\boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size *N*	*M* (number of target members)	Beta-binomial[4]	$n = N, \alpha,\ \beta$	$\alpha + \sum_{i=1}^n x_i,\ \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	$\alpha,\ \beta$	$\alpha + n,\ \beta + \sum_{i=1}^n x_i$	$\alpha - 1$ experiments, $\beta - 1$ total failures[note 1]	

Conjugate priors for continuous random variables

Wikipedia

Continuous likelihood distributions [edi

Note: In all cases below, the data is assumed to consist of n points $x_1, \ldots, x_n$ (which will be random vectors in the multivariate cases).

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive[note 4]
Normal with known variance σ^2	μ (mean)	Normal	μ_0, σ_0^2	$\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right)\Big/\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean $\bar{x}$	$\mathcal{N}(\tilde{x}\mid\mu_0', \sigma_0^{2\prime}+\sigma^2)$[5]
Normal with known precision τ	μ (mean)	Normal	μ_0, τ_0	$\left(\tau_0\mu_0+\tau\sum_{i=1}^n x_i\right)\Big/(\tau_0+n\tau), \tau_0+n\tau$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean $\bar{x}$	$\mathcal{N}\left(\tilde{x}\mid\mu_0', \left(\frac{1}{\tau_0'}+\frac{1}{\tau}\right)^{-1}\right)$[5]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	α, β [note 5]	$\alpha+\frac{n}{2}, \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance $\frac{\beta}{\alpha}$ (i.e. with sum of squared deviations 2β)	$t_{2\alpha'}(\tilde{x}\mid\mu, \sigma^2=\beta'/\alpha')$[5]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	ν, σ_0^2	$\nu+n, \frac{\nu\sigma_0^2+\sum_{i=1}^n (x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x}\mid\mu, \sigma_0^{2\prime})$[5]
Normal with known mean μ	τ (precision)	Gamma	α, β[note 3]	$\alpha+\frac{n}{2}, \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance $\frac{\beta}{\alpha}$ (i.e. with sum of squared deviations 2β)	$t_{2\alpha'}(\tilde{x}\mid\mu, \sigma^2=\beta'/\alpha')$[5]
Normal	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n}, \nu+n, \alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean $\bar{x}$; variance was estimated from $2\alpha+1$ observations with sample mean $\bar{x}$ and sample variance $\frac{\beta}{\alpha}$ (i.e. with sum of squared deviations 2β)	$t_{2\alpha'}\left(\tilde{x}\mid\mu', \frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[5]
Normal	μ and τ Assuming exchangeability	Normal-gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n}, \nu+n, \alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean $\bar{x}$, and precision was estimated from $2\alpha+1$ observations with sample mean $\bar{x}$ and sample variance $\frac{\beta}{\alpha}$ (i.e. with sum of squared deviations 2β)	$t_{2\alpha'}\left(\tilde{x}\mid\mu', \frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[5]
Multivariate normal with known covariance matrix $\boldsymbol{\Sigma}$	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$	$\left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\right), \left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}$ • $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol{\Sigma}_0^{-1}$ and with sample mean $\bar{\boldsymbol{x}}$	$\mathcal{N}(\tilde{\mathbf{x}}\mid\boldsymbol{\mu}_0', \boldsymbol{\Sigma}_0'+\boldsymbol{\Sigma})$[5]
Multivariate normal with known precision matrix $\boldsymbol{\Lambda}$	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0$	$\left(\boldsymbol{\Lambda}_0+n\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0+n\boldsymbol{\Lambda}\bar{\mathbf{x}}\right), \left(\boldsymbol{\Lambda}_0+n\boldsymbol{\Lambda}\right)$ • $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol{\Lambda}$ and with sample mean $\bar{\boldsymbol{x}}$	$\mathcal{N}\left(\tilde{\mathbf{x}}\mid\boldsymbol{\mu}_0', (\boldsymbol{\Lambda}_0'^{-1}+\boldsymbol{\Lambda}^{-1})^{-1}\right)$[5]
Multivariate normal with						

ebayes{limma}

- Gordon Smyth, (2004). **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. Statistical Applications in Genetics and Molecular Biology, Volume **3**
- empirical Bayes shrinkage of the standard errors towards a common value
- borrow information from all genes to infer the variance for each group of replicates

```
#  Simulate gene expression data,
#  6 microarrays and 100 genes with one gene differentially expressed
set.seed(2004); invisible(runif(100))
M <- matrix(rnorm(100*6,sd=0.3),100,6)
M[1,] <- M[1,] + 1
fit <- lmFit(M)

#  Moderated t-statistic
fit <- eBayes(fit)
topTable(fit)
```

https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/ebayes