I know what you ate last summer!

- … with some uncertainty, of course… …or not.

- *Outline:*

➢Practical use of Bayesian statistics for simple problems. Example.

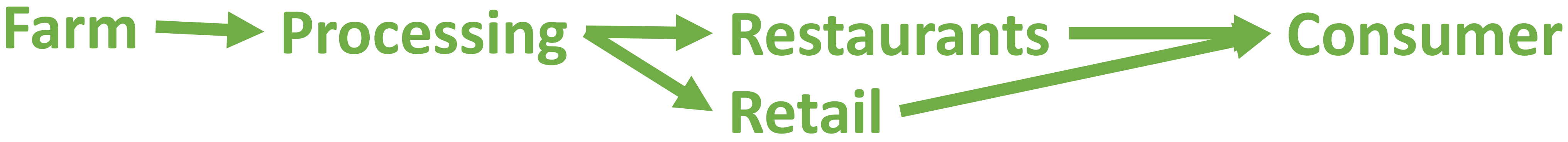
➢Bayes for evidence synthesis.

➢Bayes for source attribution.

➢Bayes for acute food consumption risk and prediction.

Bayes for risk assessment in food safety

- Food safety depends on lots of things from farm-to-fork



Farm

Processing

Restaurants

Retail

Consumer

- Not enough to *’know what you typically eat’*, but also:
    - how much/often you eat,
    - how you made/kept it,
    - where did you buy it,
    - and how it was produced !
- Some data from these steps.
- Bayesian methods exploited to quantify probabilities.

- More Bayesian Food Safety Risk Assessment applications

ELSEVIER

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

Hierarchical Bayesian Modelling for *Saccharomyces cerevisiae* population dynamics

Aymé Spor [a,*], Christine Dillmann [a], Shaoxiao Wang [a,b], Dominique de Vienne [a], Delphine Sicard [a], Eric Parent [c]

[a] Univ Paris-Sud, UMR 0320/UMR 8120 Génétique Végétale, F-91190 Gif-sur-Yvette, France
[b] Department of Biochemistry and Molecular Biology, Louisiana State University Health Sciences Center, Shreveport, LA 71130, USA
[c] UMR MIA 518, INRA/AgroParisTech, ENGREF, 19 avenue du Maine, F-75015 Paris, France

ARTICLE INFO

Article history:

ABSTRACT

Hierarchical Bayesian Modelling is powerful however under-used to model and evaluate the risks associated with the development of pathogens in food industry, to predict exotic invasions, species extinctions and development of emerging diseases, or to assess chemical risks. Modelling population dynamics of *Saccharomyces cerevisiae* considering its biodiversity and other sources of variability is crucial for selecting strains meeting industrial needs. Using this approach, we studied the population dynamics of *S. cerevisiae*, the domesticated yeast, widely encountered in food industry, notably in brewery, vinery, bakery and distillery. We relied on a logistic equation to estimate the key variables of population growth, but we took also into account factors able to affect them, namely environmental effects, genetic diversity and measurement errors. Our probabilistic approach allowed us: (i) to model the dynamical behaviour of

Risk Analysis, Vol. 28, No. 2, 2008

DOI: 10.1111/j.1539-6924.2008.01000.x

Quantitative Risk Assessment from Farm to Fork and Beyond: A Global Bayesian Approach Concerning Food-Borne Diseases

Isabelle Albert,[1] Emmanuel Grenier,[2] Jean-Baptiste Denis,[3] and Judith Rousseau[4]

A novel approach to the quantitative assessment of food-borne risks is proposed. The basic idea is to use Bayesian techniques in two distinct steps: first by constructing a stochastic *core model* via a Bayesian network based on expert knowledge, and second, using the *data* available to improve this knowledge. Unlike the *Monte Carlo simulation* approach as commonly used in quantitative assessment of food-borne risks where data sets are used independently in each module, our consistent procedure incorporates information conveyed by data throughout the chain. It allows "back-calculation" in the food chain model, to
data obtained "downstream" in the food chain. Moreover, the expert kr
more simply and consistently than with classical statistical methods. Ot
approach include the clear framework of an iterative learning process
ity enabling the use of heterogeneous data, and a justified method to
variability and uncertainty. As an illustration, we present an estimatio
contracting a campylobacteriosis as a result of broiler contamination, f
quantitative risk assessment. Although the model thus constructed is ov
the principles and properties of the method proposed, which demonst
with quite complex situations and provides a useful basis for further dis
experts in the food chain.

KEY WORDS: Bayesian network; Bayesian statistics; broiler; campylobacteri
risk assessment

1. INTRODUCTION

Quantitative risk assessment (QRA) regarding food-borne pathogens is growing in importance, for both public health and trade purposes. International organizations such as the WHO and FAO, and vari-

needs to be populated with valı
approach called process risk mc
modular process risk modeling (
splits the farm-to-fork chain intc
cally and sequentially progress i
that of the food system. The sta

Risk Analysis, Vol. 31, No. 7, 2011

DOI: 10.1111/j.1539-6924.2010.01572.x

A Bayesian Evidence Synthesis for Estimating Campylobacteriosis Prevalence

Isabelle Albert,[1,*] Emmanuelle Espié,[2] Henriette de Valk,[2] and Jean-Baptiste Denis[3]

Risk Analysis, Vol. 33, No. 5, 2013

DOI: 10.1111/j.1539-6924.2012.01888.x

Inferring an Augmented Bayesian Network to Confront a Complex Quantitative Microbial Risk Assessment Model with Durability Studies: Application to *Bacillus Cereus* on a Courgette Purée Production Chain

Clémence Rigaux,[1,*] Sophie Ancelet,[1] Frédéric Carlin,[2,3] Christophe Nguyen-thé,[2,3] and Isabelle Albert[1]

1

Practical usefullness: doing simple statistics

Often small sample analyses are done using various statistical tests.

Worries:

- What test to use?
- Is the sample size large enough?
- Is the number of blocks/groups large enough?
- Interpretation of results: reject H_0 or not, and what to say then?
- Multiple testing problems…
- Testing just because of the habit?

- Biofilm production by 10 strains of S.Enteritidis on cutting boards

Material	None	Weak	Moderate	Strong
Wood	4	5	1	0
Plastic	6	4	0	0
Glass	9	1	0	0

- What are we really asking?
  - Which material is safest?

- How does it translate to a statistical question?
  - Q1: do the materials differ?
  - Q2: which material has the highest P(None)?

[Foodborne pathogens and disease 15, (2), 2018. 81-85.]

Bayesian formulation of the problem

Model: Multinomial probabilities

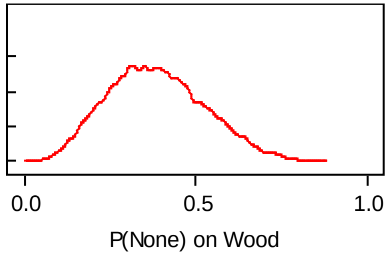
p_1, p_2, p_3, p_4 =

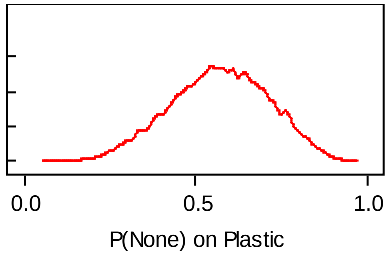
P(None), P(Weak), P(Moderate), P(Strong)

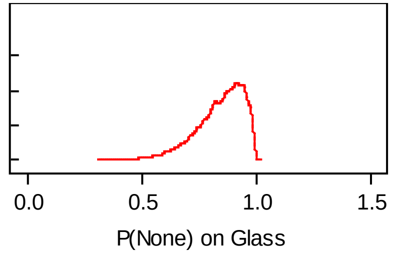
Compute P(p_1, p_2, p_3, p_4 | data) for each material

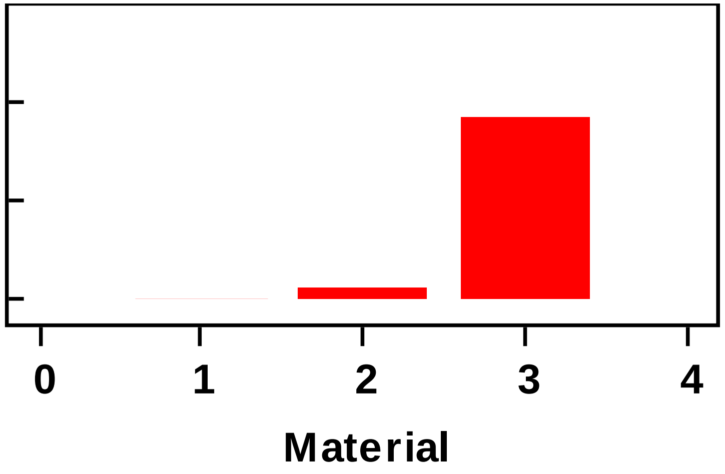
- Typical prior: Dirichlet(1/4,…,1/4)
- →Posterior: Dirichlet(x_1+1/4,…,x_4+1/4)
- All conclusions produced from this!

- For example:
  - P(P(None) is highest on Glass)









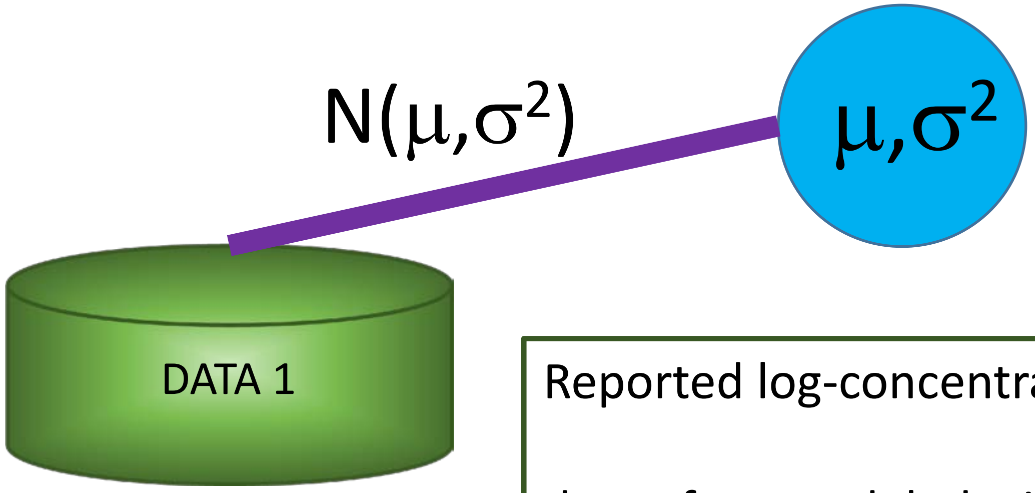
Take home recipe: simple-to-run code for OpenBUGS/WinBUGS

```
model{
for(i in 1:materials){
 pnone[i] <- p[i,1]
 p[i,1:k] ~ ddirch(a[i,1:k])
 for(j in 1:k){a[i,j] <- x[i,j]+1/k}
}
largest.value <- ranked(pnone[],materials)
for(i in 1:materials){ which[i] <- equals(pnone[i],largest.value)*i }
pnonelargest <- sum(which[])
}
# data:
list(materials=3,k=4,
x=structure(.Data=c(
4,5,1,0,
6,4,0,0,
9,1,0,0),.Dim=c(3,4)))
```

Simple Bayesian models for simple problems can also be useful, and not too hard to implement.

2

Simple evidence synthesis: $N(\mu,\sigma^2)$



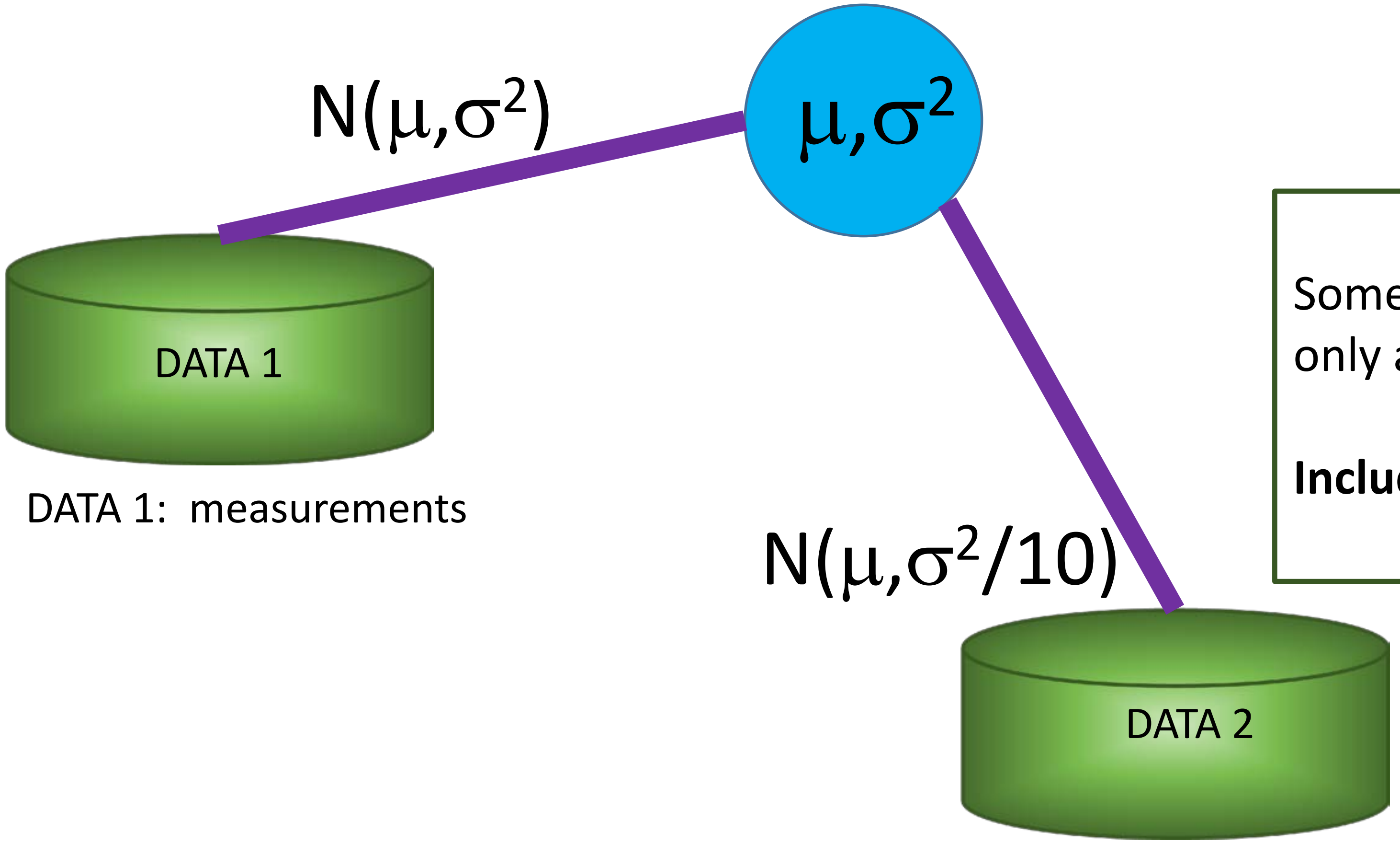
DATA 1: measurements

Reported log-concentrations:

data often modeled with parametric distributions, e.g. normal.

DATA1: this goes in easily!

Simple evidence synthesis: $N(\mu,\sigma^2)$



DATA 1: measurements

DATA 2: averages of 10 measurements

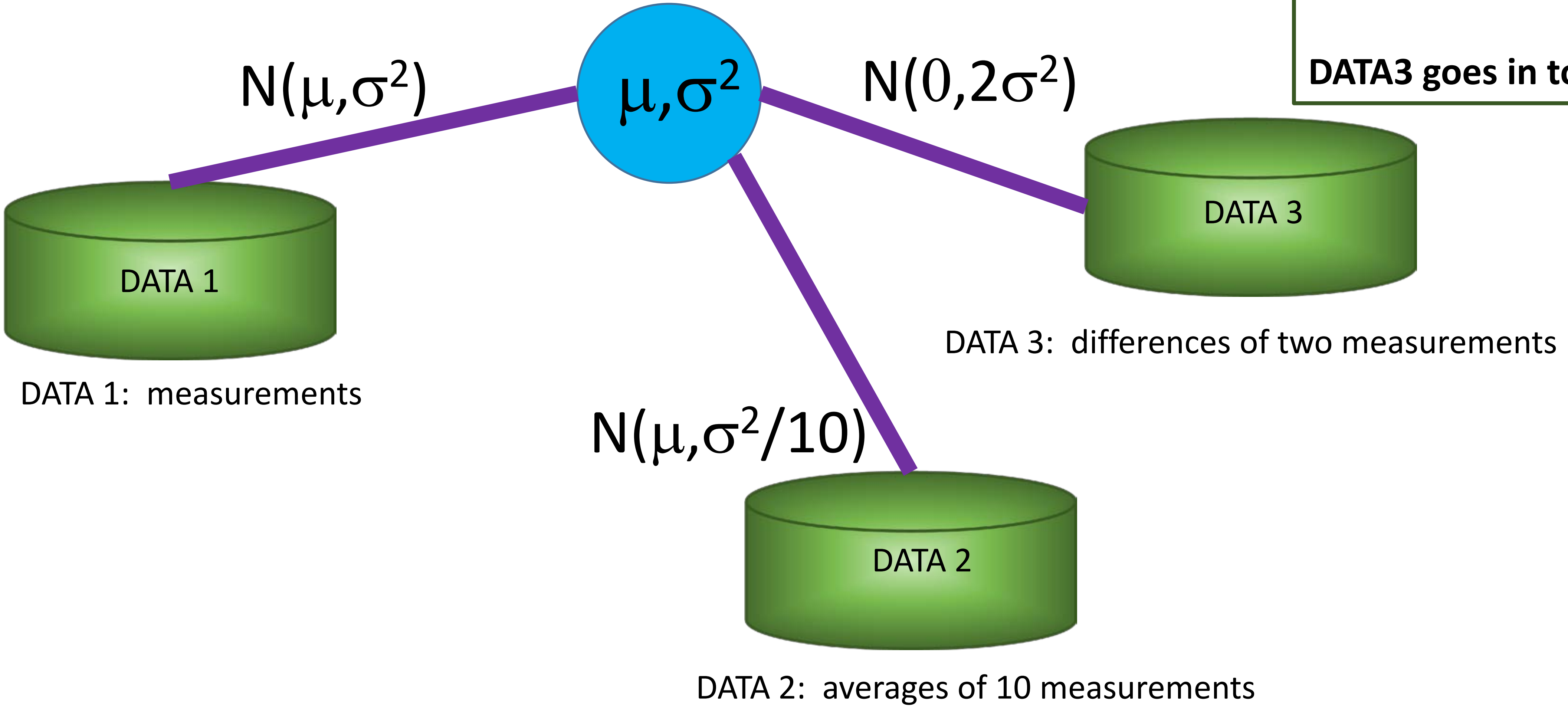
Some data could be reported only as averages

Include also DATA2

Simple evidence synthesis: N(μ,σ^2)

Or reported differences

DATA3 goes in too!



μ,σ^2

N(μ,σ^2)

DATA 1

DATA 1: measurements

N(μ,σ^2/10)

DATA 2

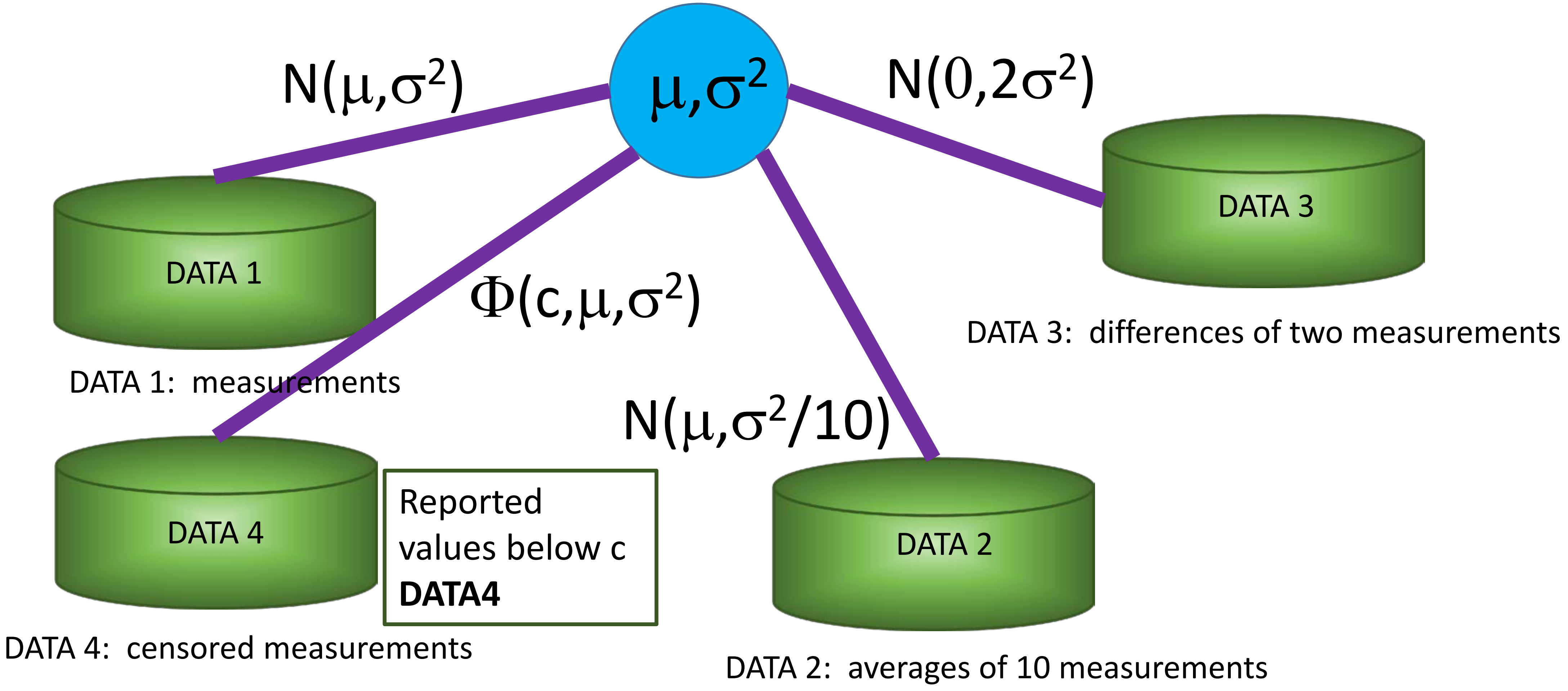
DATA 2: averages of 10 measurements

N(0,2σ^2)

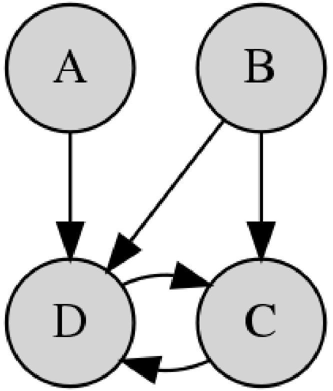
DATA 3

DATA 3: differences of two measurements

Simple evidence synthesis: $N(\mu,\sigma^2)$


μ,σ^2
$N(\mu,\sigma^2)$
$N(0,2\sigma^2)$
$\Phi(c,\mu,\sigma^2)$
$N(\mu,\sigma^2/10)$
DATA 1
DATA 3
DATA 4
DATA 2
DATA 3: differences of two measurements
DATA 1: measurements
Reported values below c **DATA4**
DATA 4: censored measurements
DATA 2: averages of 10 measurements

If there is a model, there's a way



Maximum likelihood estimation

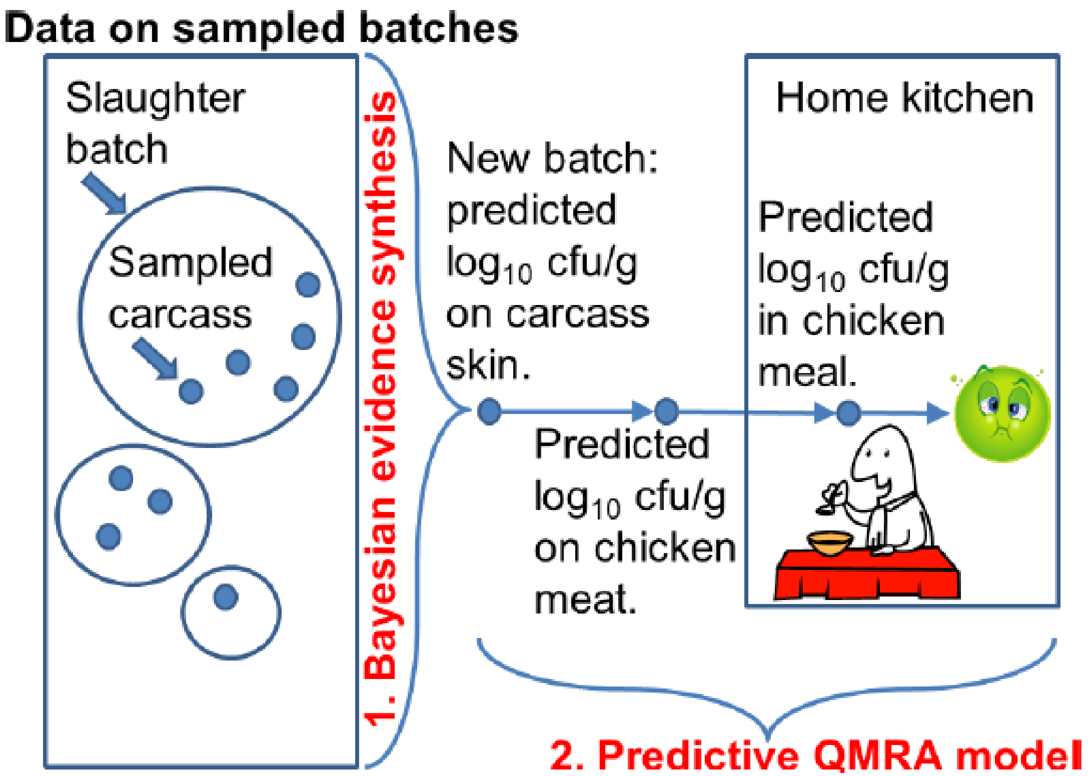
- Construct full likelihood of all datasets.
- Maximise to get ML-estimates
- Higher dimensions can become difficult.
- Multiple maxima?
- Aiming to get **the single estimate**.

Bayesian inference

- Construct full likelihood of all datasets.
- Define prior distributions.
- Simulate the posterior distribution using MCMC (BUGS,JAGS,STAN,own sampler).
- Aiming to get **the uncertainty distribution** of all parameters.

Is there Campylobacter in the broiler you get?

- Your broilers are ’sampled’ from production batches.
- There is variability between batches and within batches.


Data on sampled batches
Slaughter batch
Sampled carcass
1. Bayesian evidence synthesis
New batch: predicted $\log_{10}$ cfu/g on carcass skin.
Predicted $\log_{10}$ cfu/g on chicken meat.
Home kitchen
Predicted $\log_{10}$ cfu/g in chicken meal.
2. Predictive QMRA model

→ consumers’ risk

Do we have enough evidence for an estimate?

- There were two (Swedish) data sets:
  - A: representing only one broiler from each batch, 10 batches sampled in a representative way. Result: positive/negative, & concentration if positive.

    → 88 pos, 617 neg, hence 88 conc. values.
  - B: representing the mean and SD of log-concentrations, from 5 to 25 positive broilers per batch, from 20 positive batches, and the # posit/negat broilers in each batch.

Complementing evidence from both

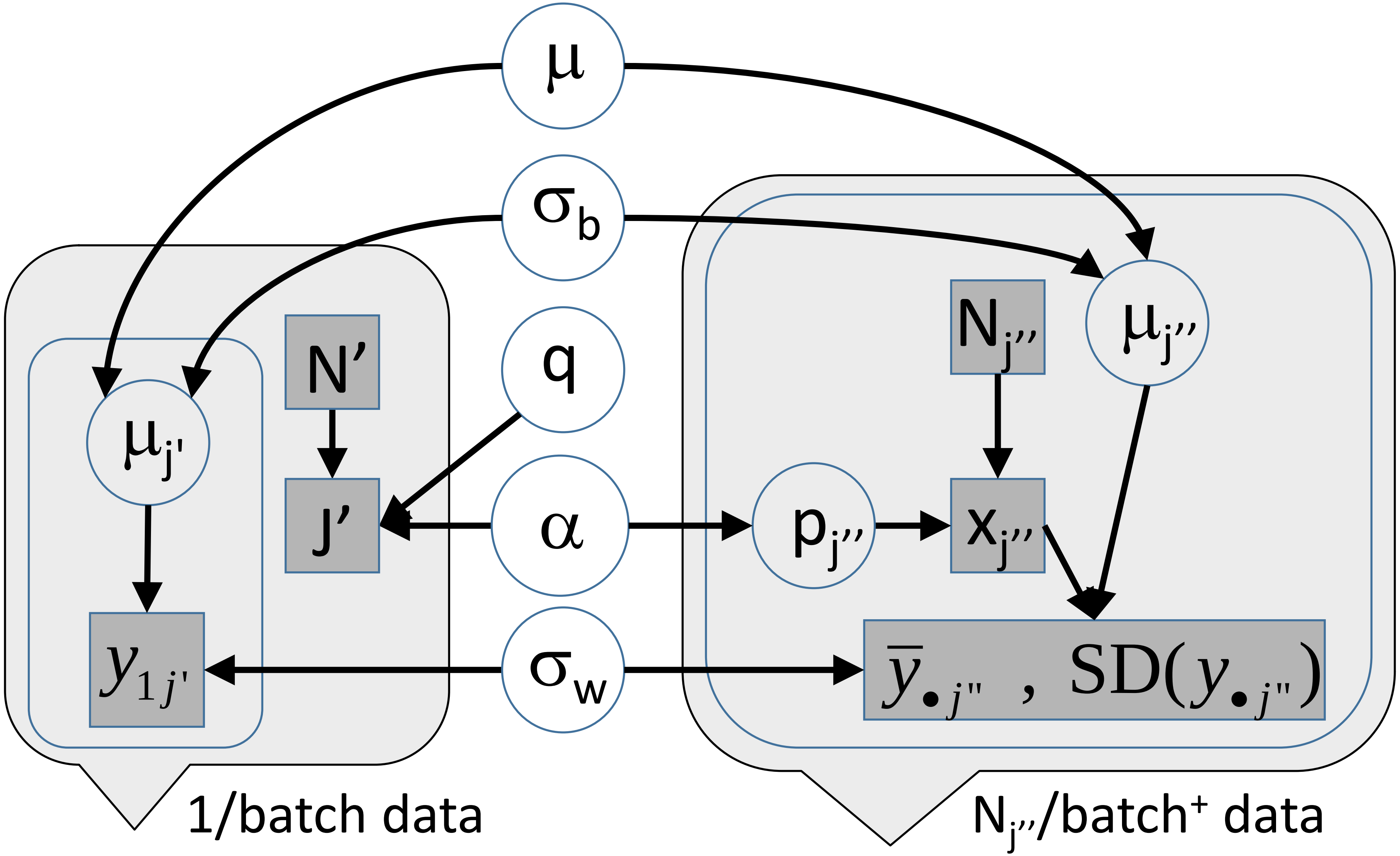
- **A:** information about mean and total variance of concentrations in positive broilers, but nothing about within-batch prevalence*, or variance components.

  (*) if we assume within-batch prevalence 100%, can estimate batch prevalence.

- **B:** information on within-batch parameters for positive batches, but nothing on overall batch prevalence.

- **Make a synthesis of A & B with a Bayesian model.**

Just like in the example before: models connected with common parameters



Posterior distributions for the two variance components

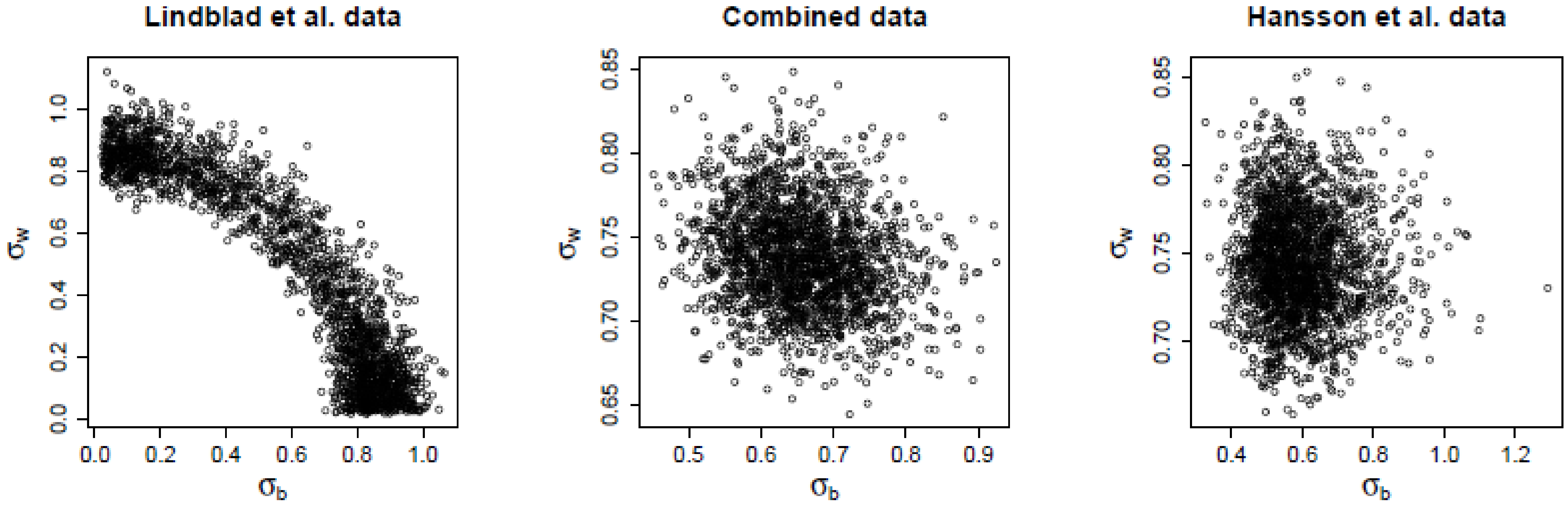
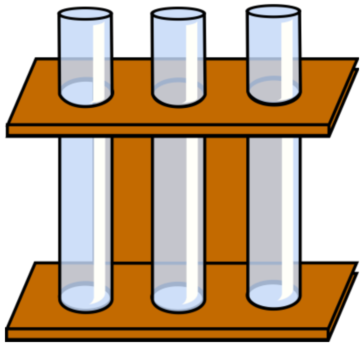


FIG 3. *Marginal posterior distributions of* (σ_b, σ_w) *based on each data set alone (1/batch left,* $N_{j''}$*/batch*$^+$ *right) and the two data sets combined (middle).*

Estimation from a synthesis is *interesting*, but there's more than that…



- A **M**icrobiological **C**riterion (**MC**) can be placed for the acceptance of a batch.
- This would be based on sampling results, batch by batch.
- When bad batches are rejected, consumers' risk is reduced.
- But producers' costs are increased if too many batches are rejected!

consumers' risk

VS

producers' risk

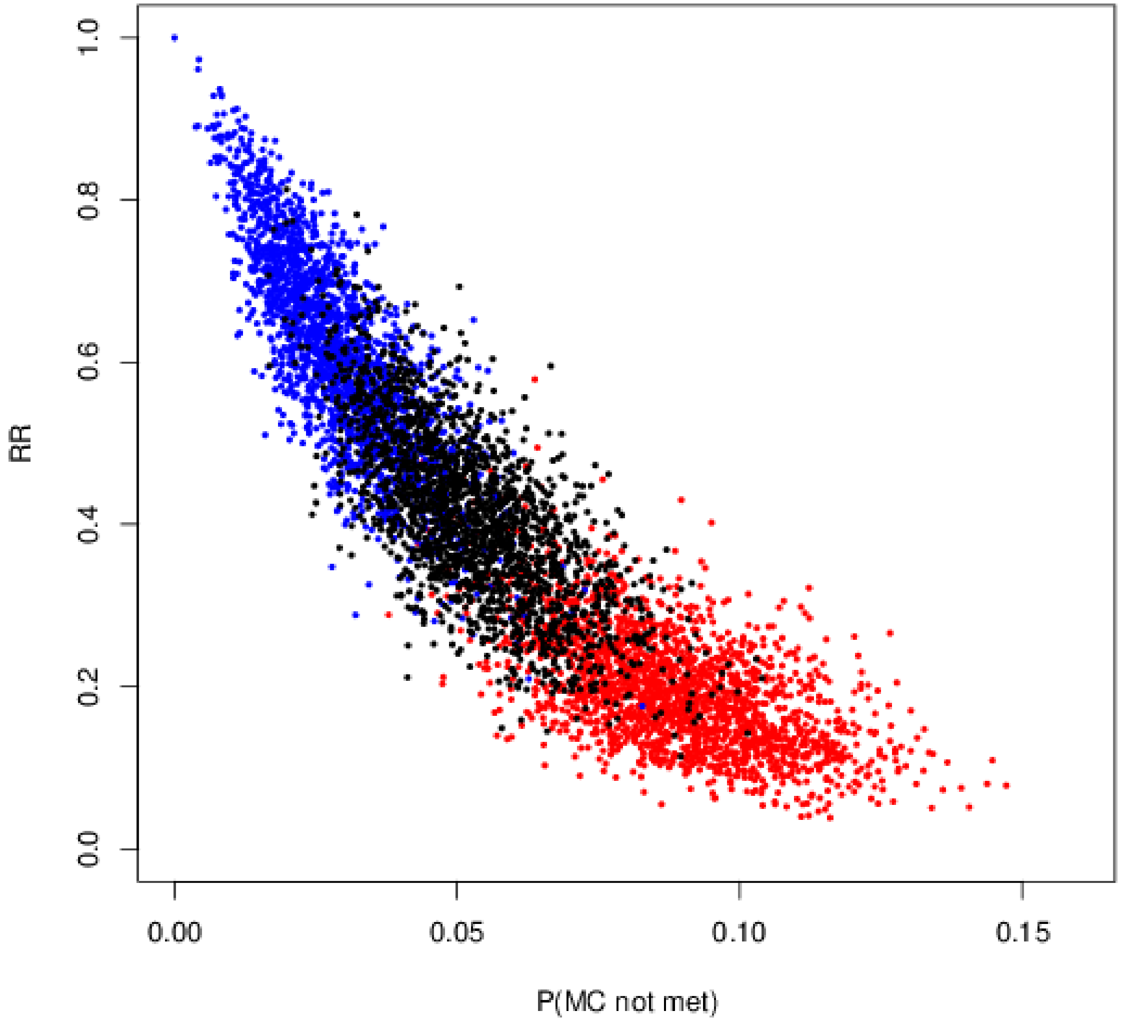
What does the outcome from such test sample represent? - Additional evidence.

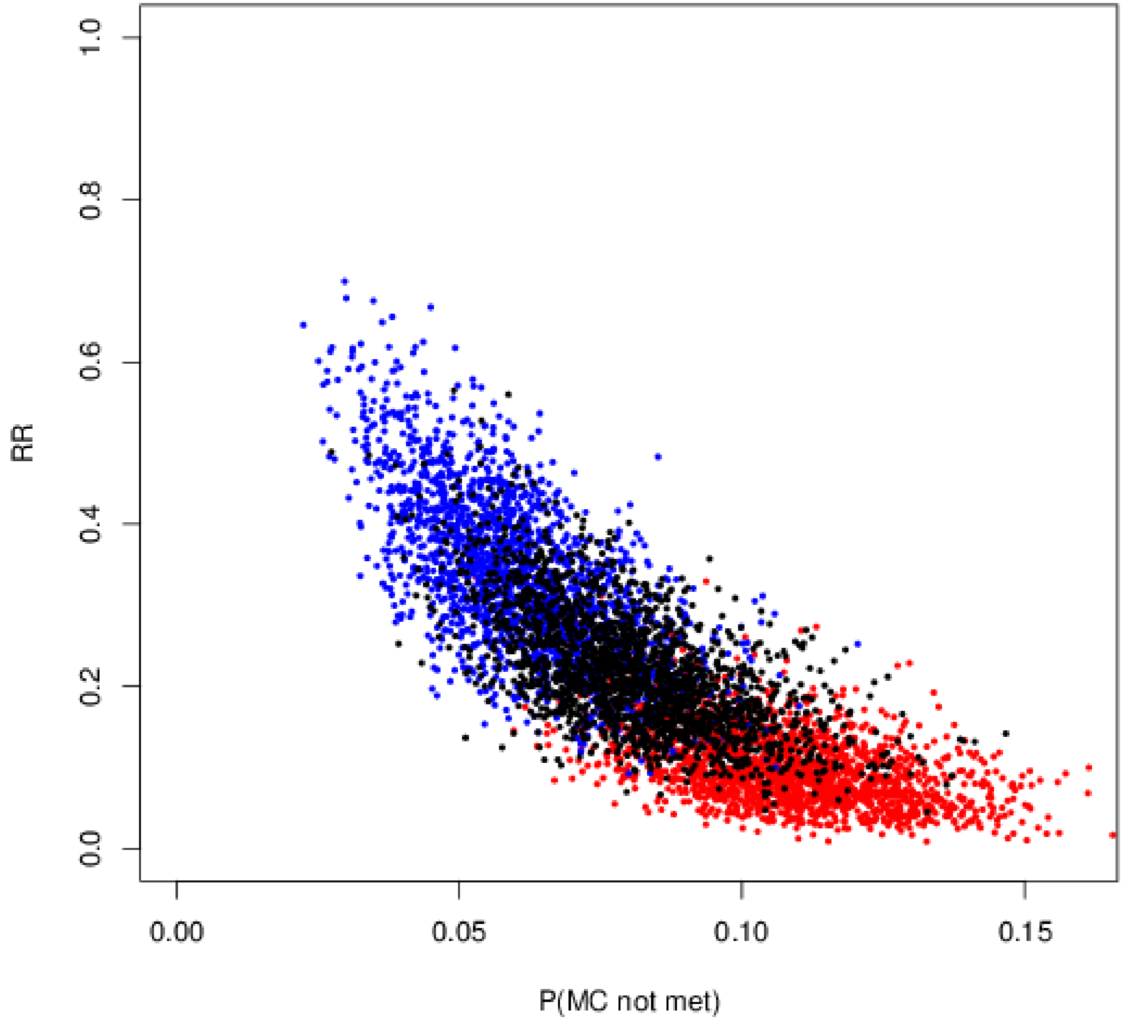
- Can use Bayesian model to revise the estimates for PREDICTED ACCEPTED batches.
- This determines the new consumer risk, under such criterion.
- Can also calculate the probability of rejection for batches → predicted percentage of lost batches.
- A criterion could be: ”n/c/m” = ”at most c samples out of n are allowed to have concentration >m”.
- **HOW TO CHOOSE n/c/m ?**
- Uncertainty analysis involves 2D Monte Carlo (MC within MCMC).

Finding an optimal criterion, accounting for uncertainties.

RR = risk ratio = risk when MC is met / risk if no MC was applied.

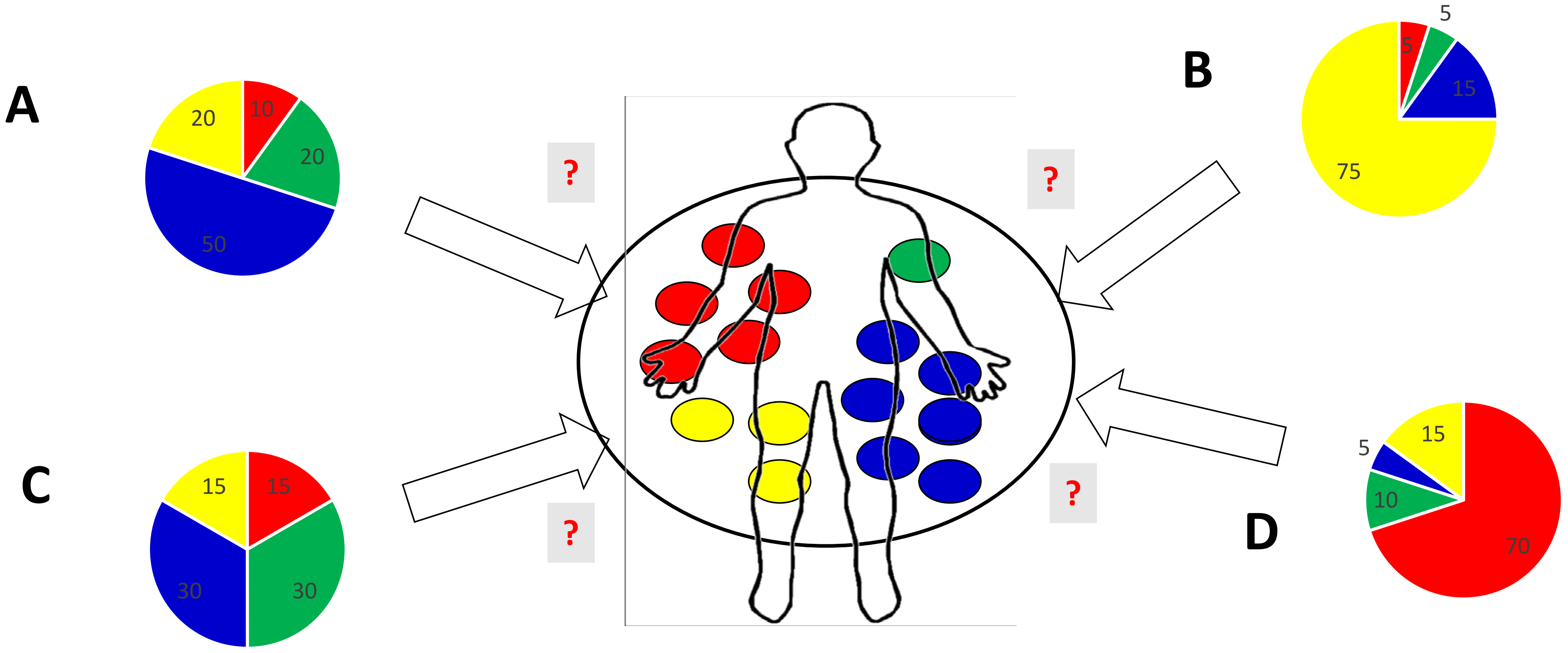
P(MC not met) = percentage of rejected batches.



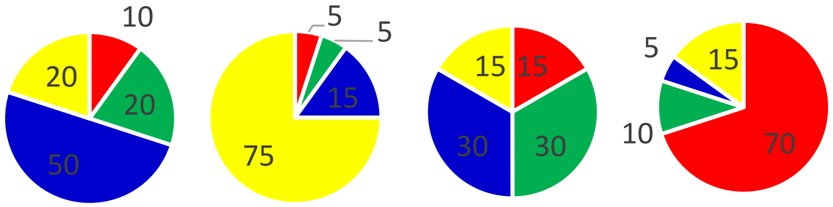


3

Classification problems: ’source attribution’



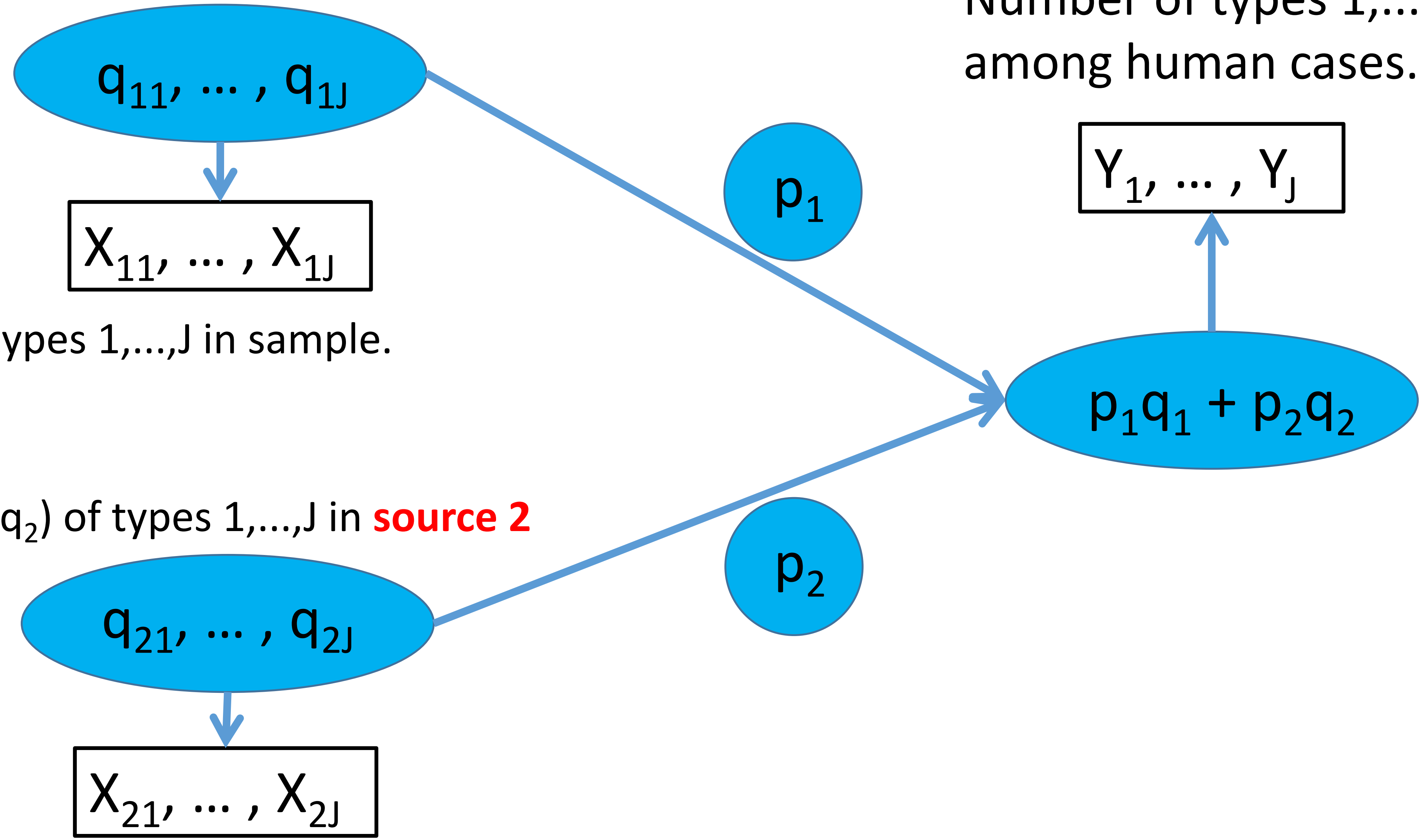
- **Bacteria types sampled from a few broad food categories, denoted as ’the sources’.**



  - E.g. broilers (samples from meat and/or animals),
  - Likewise turkey , cattle, pigs, etc.
  - Possibly also other exposures: swimming waters, environment,…

- Bacteria types from human isolates taken as **a mixture sample of sources**.

- **Problem:** assuming human isolates (somehow) originated from those sources,
  - classify each isolate into sources.
  - estimate what fraction of cases are generally from which source (mixture proportions).



Proportion (q_1) of types 1,…,J in **source 1**

$q_{11}, \dots , q_{1J}$

$X_{11}, \dots , X_{1J}$

Number of types 1,…,J in sample.

Proportion (q_2) of types 1,…,J in **source 2**

$q_{21}, \dots , q_{2J}$

$X_{21}, \dots , X_{2J}$

Number of types 1,…,J in sample.

p_1

p_2

$p_1q_1 + p_2q_2$

$Y_1, \dots , Y_J$

Number of types 1,…,J among human cases.

Bayesian classification methods

- *Naive Bayes classifier* with sources i = 1,…,I, and types j=1,…,J
  - P(source i | type j) = P(type j | source i) P(source i) / const

  - P(source i) = 1/I, prior probability, i=1,…,I sources.
  - P(type j| source i) = Multinomial(q_i*,1) with **estimated type frequencies q_i*** directly from data: q_{ij}* = x_{ij} / n_i or smoothed: (x_{ij} +1/J)/ (n_i+1).

  - If P(source i) = p_i with prior P(p_i), we obtain **posterior distribution: P(I_1,…,I_N,p_1,…,p_I | x,y)** for the **population fractions p** (mixture proportions), *and* **source labels** I_n for each human case, based on source samples x and human samples y.

Bayesian classification methods

- *Posterior predictive classifier*
  - For a single new isolate ***in a source i***, predictive probability: P(type j |x_i) = P_{ij}

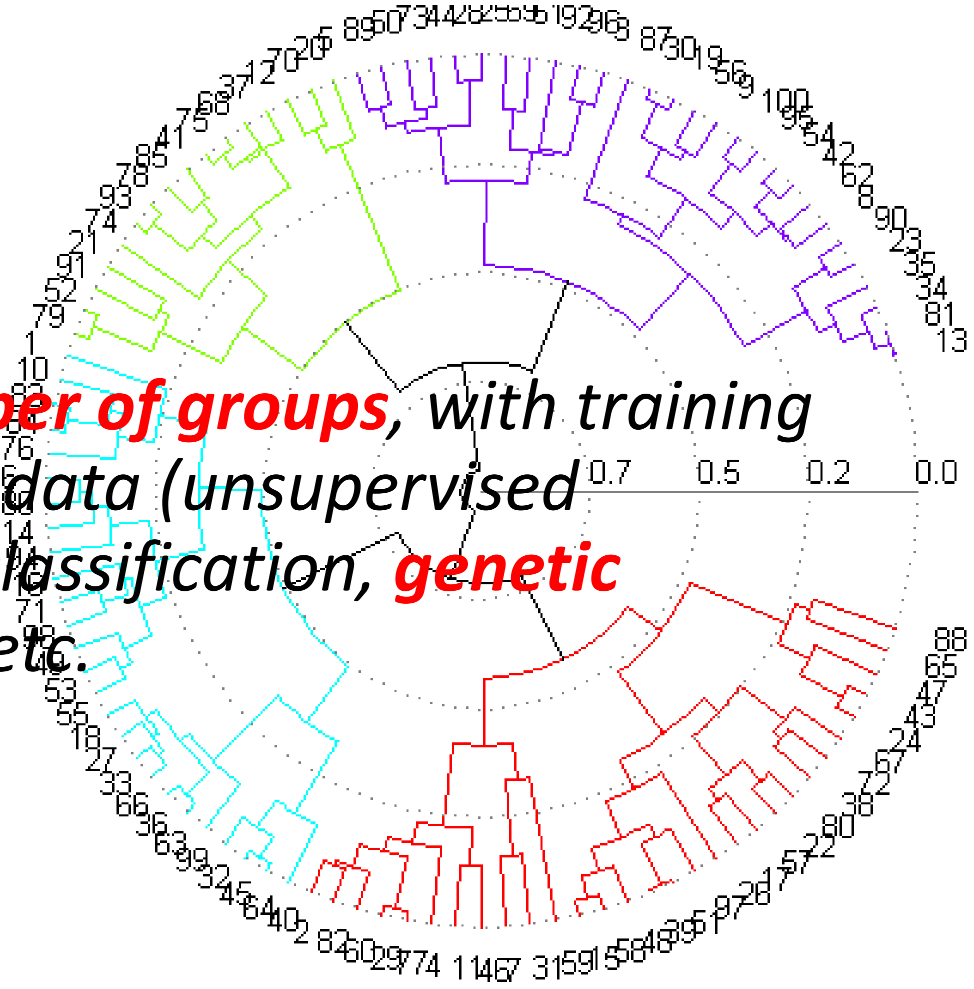
$$P_{ij} = \frac{\Gamma(\sum \alpha_{ij})}{\Gamma(1+\sum \alpha_{ij})} \frac{\Gamma(1+\alpha_{ij})}{\Gamma(\alpha_{ij})} \text{ from the integral (predictive distribution) :}$$

$$\int \text{Multin}(q_{i1},\ldots,q_{iJ},1)\text{Dir}(q_{i1},\ldots,q_{iJ} \mid \alpha_{i1},\ldots,\alpha_{iJ})Dq$$

  - α_j are parameters of the posterior distribution of the type frequencies q in that source.

  - These predictive probabilities can be used to evaluate
    P(source i | type j, $x_{1, \ldots,}$ x_I) = P(type j | x_i) P(source i) / const.

More advanced methods

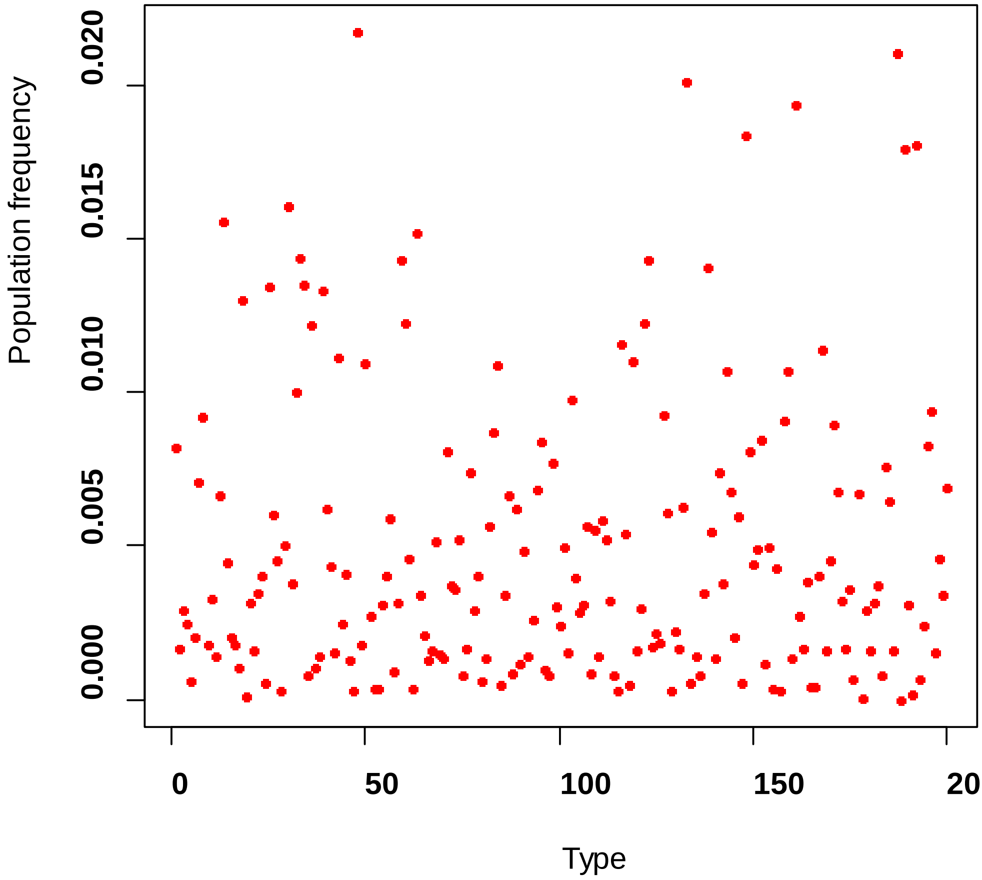


- *Extensions: clustering into* ***unknown number of groups****, with training data for some groups, or without training data (unsupervised classification), marginal or simultaneous classification,* ***genetic population structure****, evolutionary trees, etc.*

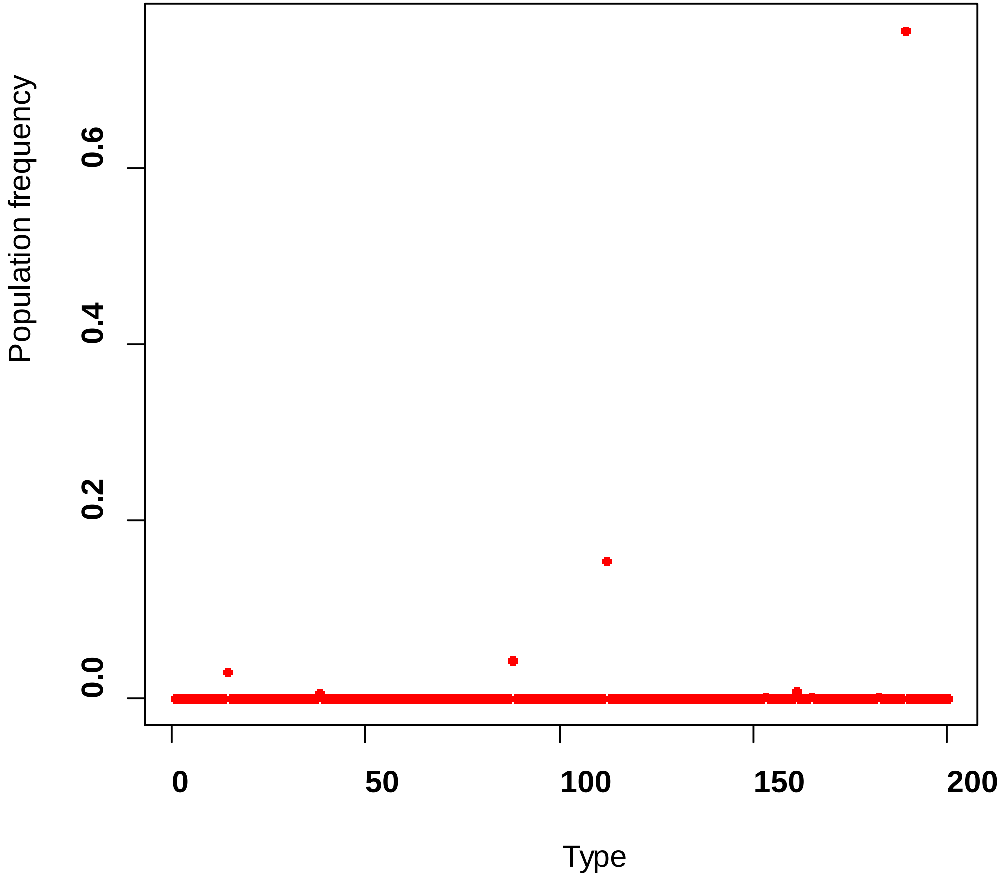
  - Here we simply restrict to a fixed defined number of groups described as ”the food product types”, each simply represented with some surveillance samples, with a set of possible types.

Simulation experiment with 200 types, 5 sources.

If many types are frequent in a source
Dirichlet(1,…,1)

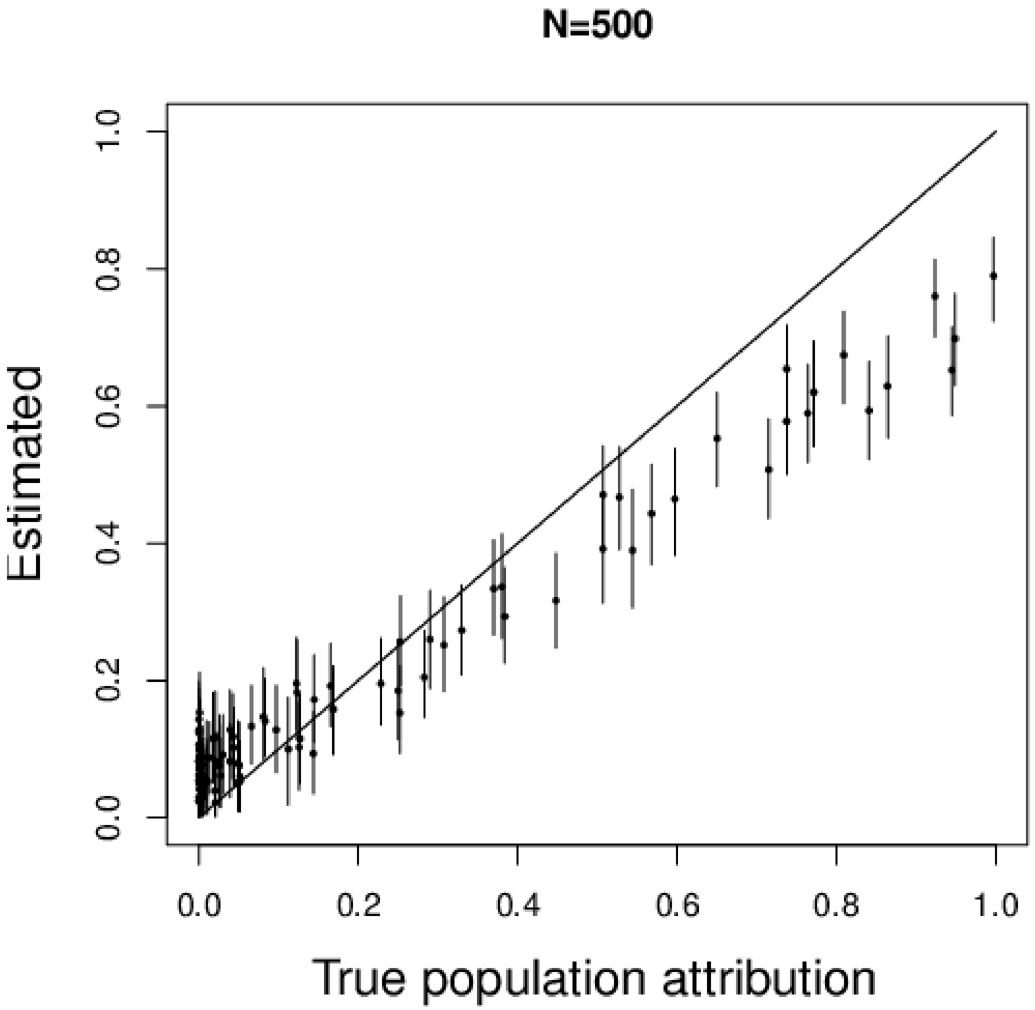


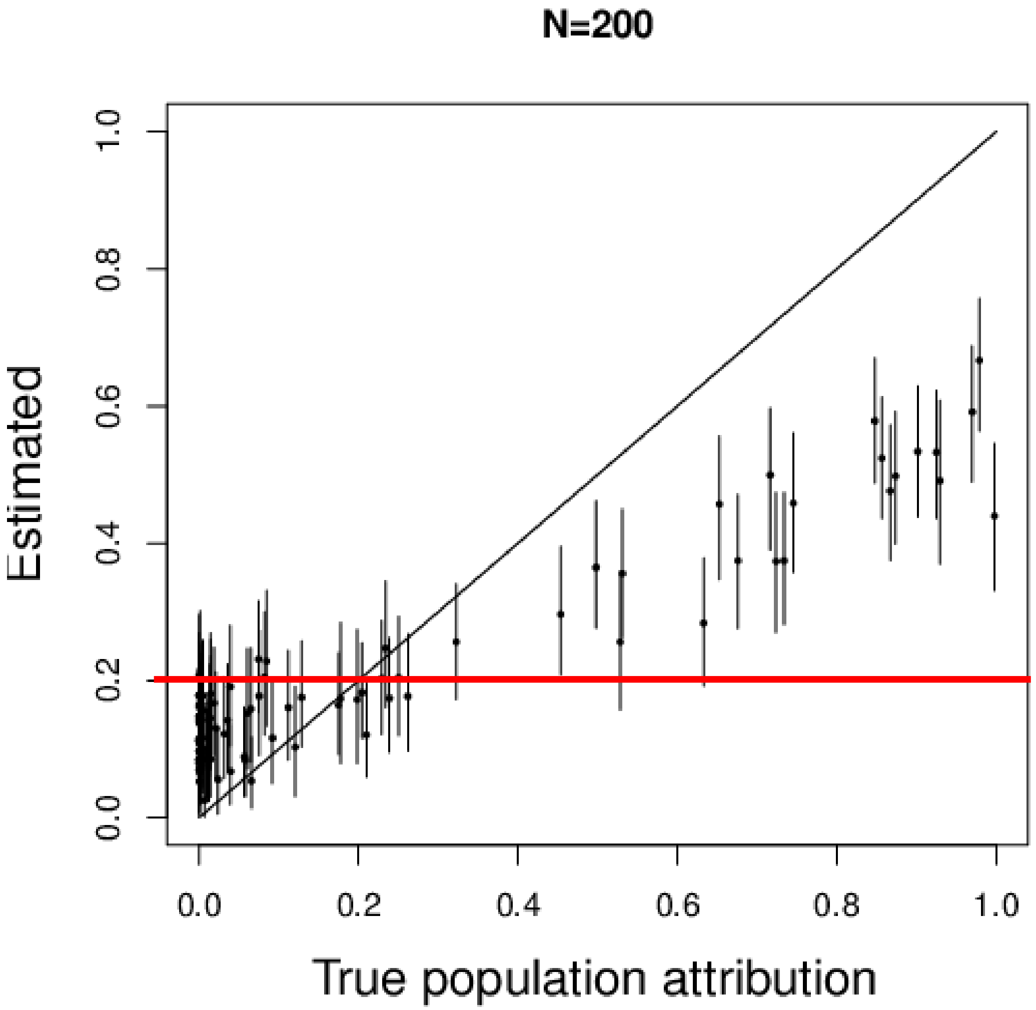
If only a few types are frequent in a source
Dirichlet(1/200,…,1/200)



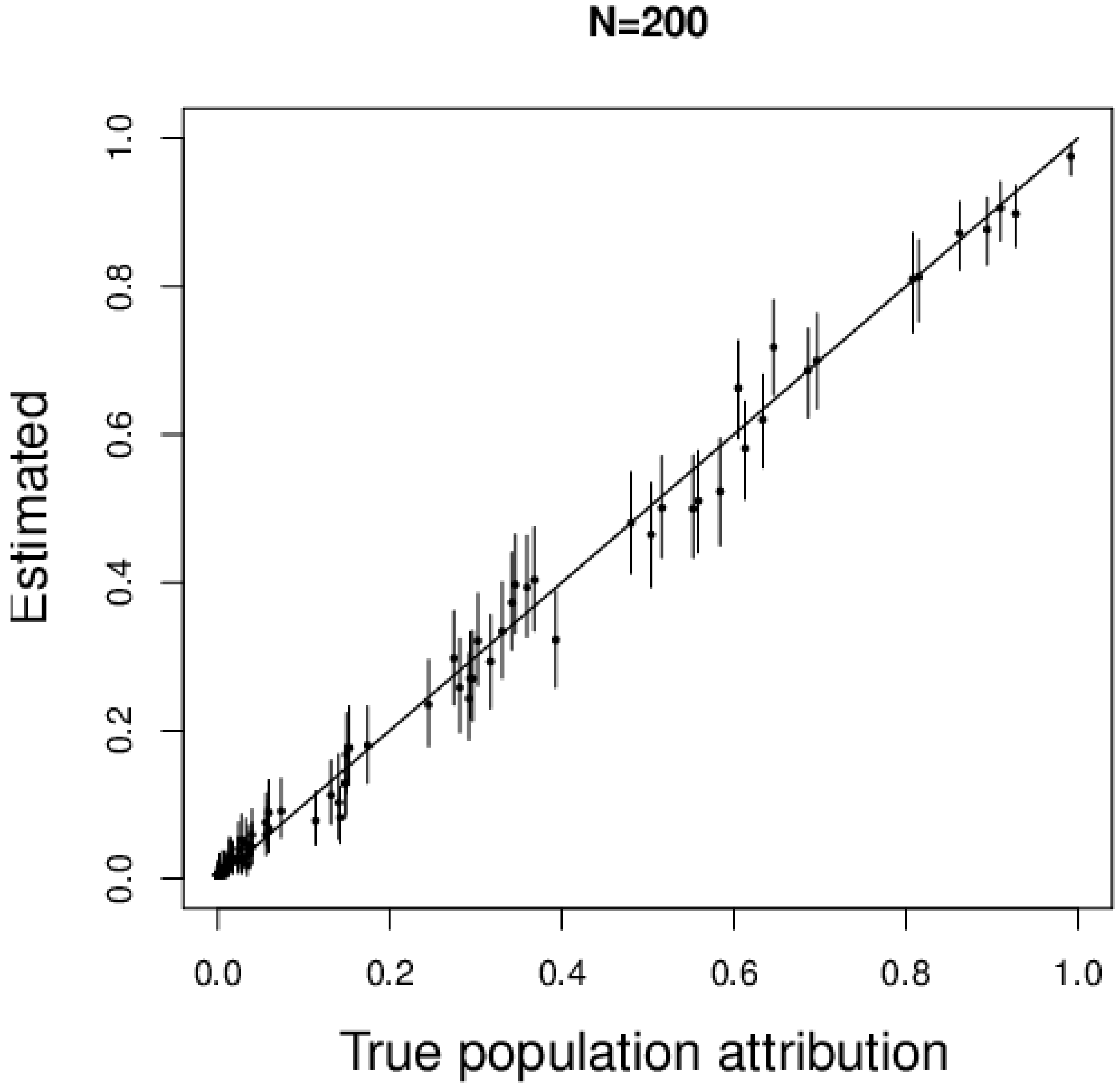
If many types are evenly frequent in each source

5 sources 5*N samples, N human cases





Easy if only a few types dominate in each source



4

Basics of intake assessment

- Acute exposure to a microbe or chemical: $\sum C_k * W_k$
  - C_k = concentration of the hazardous substance per gram, in food type k.
  - W_k = consumption amount (in grams) of food type k per serving.
  - C and W independent → a model for both occurrence data and consumption data.
  - Variability of C between samples, between food types.
  - Variability of W between days (for an individual), between individuals.

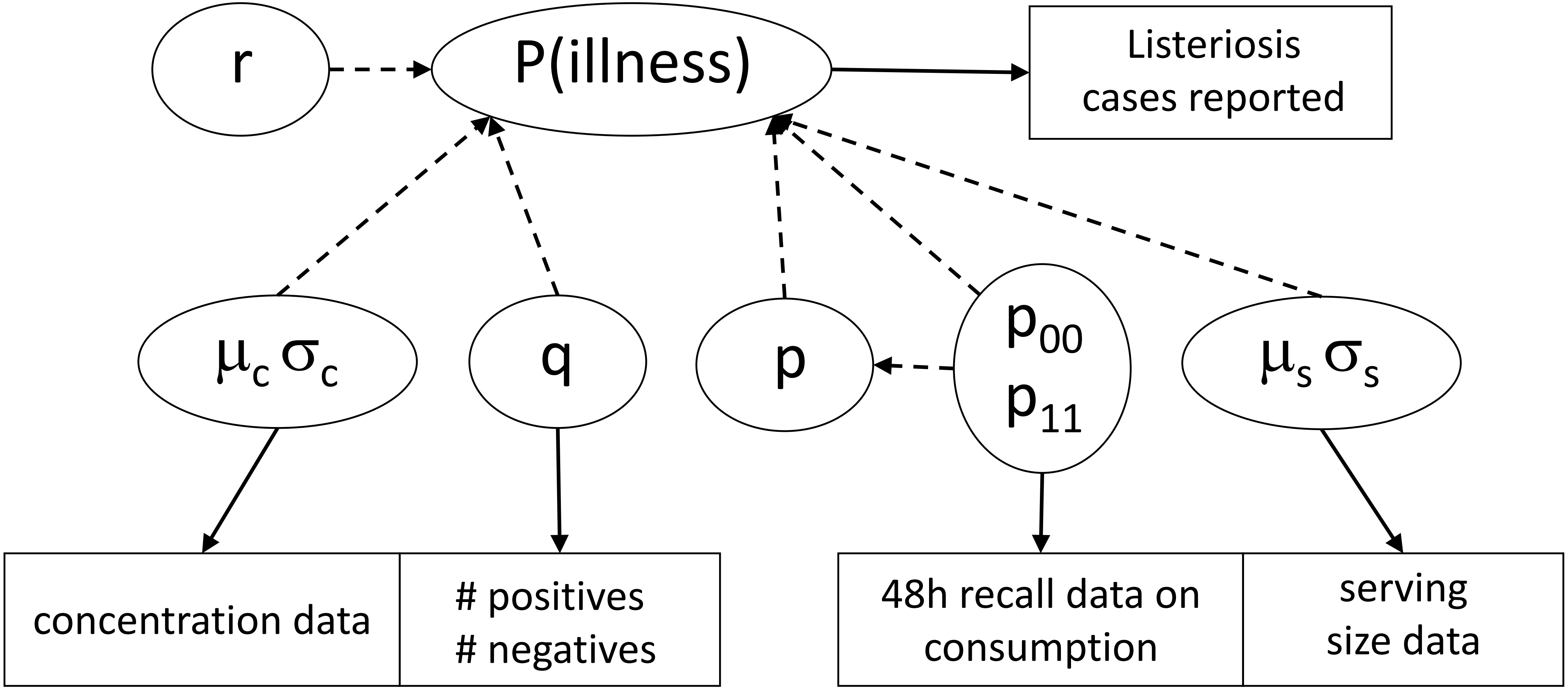
Listeria risk for a whole week and beyond!

- Acute risk of illness for e.g. Listeria from ready-to-eat food: the probability per serving (or per day) to get ill.

  - Depends on:
  - The probability to consume that food (consumption frequency).
  - The probability of Listeria in it (prevalence of Listeria).
  - The probability distribution of consumption amount.
  - The probability distribution of Listeria concentration.
  - Dose-response probability.

  - And the **growth** of bacteria during storage if you eat it later (again).
  - And how likely you would **continue** eating the following days.

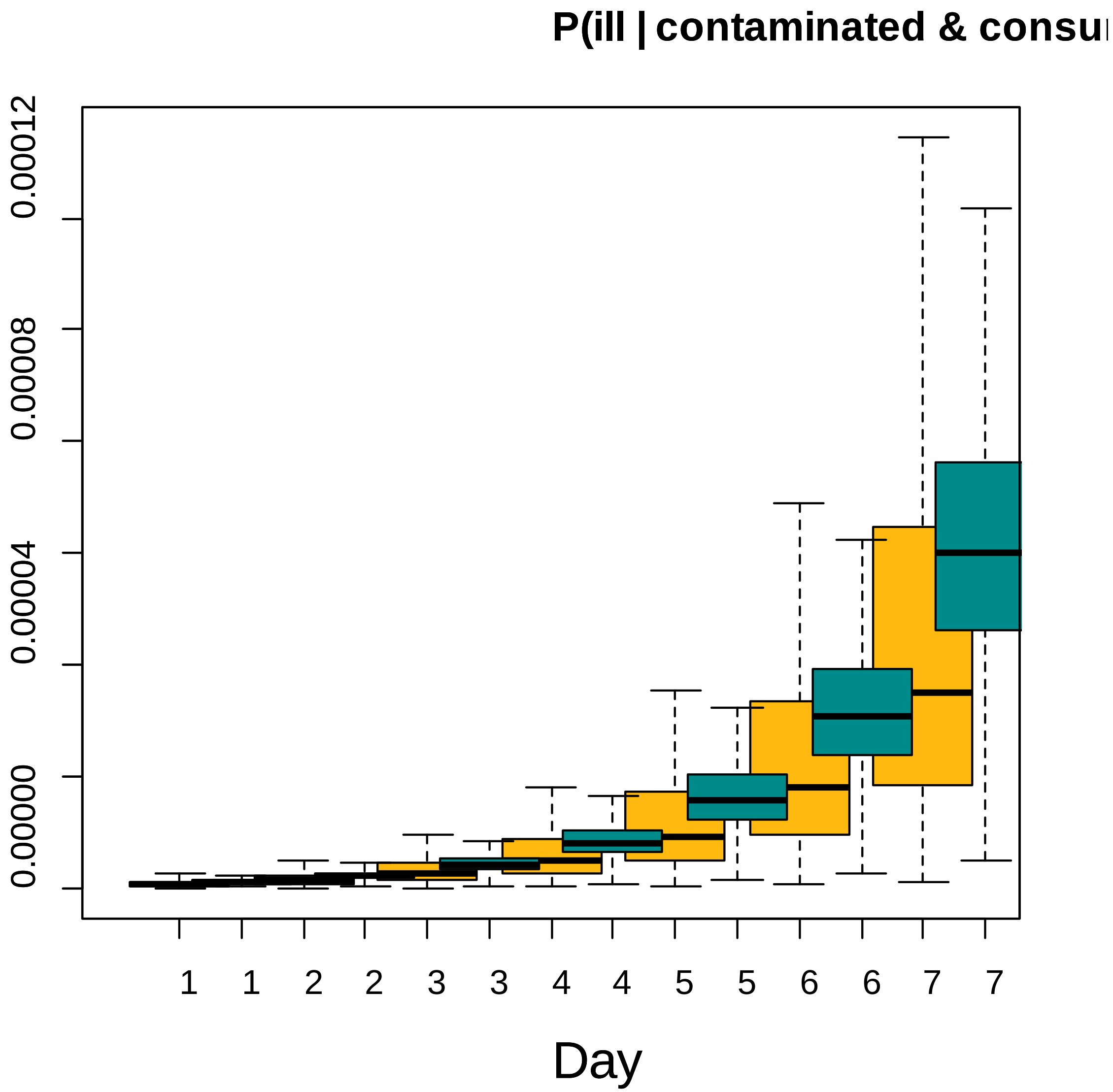
Again, start with evidence synthesis for all parameters



If you consume bad food on a bad day

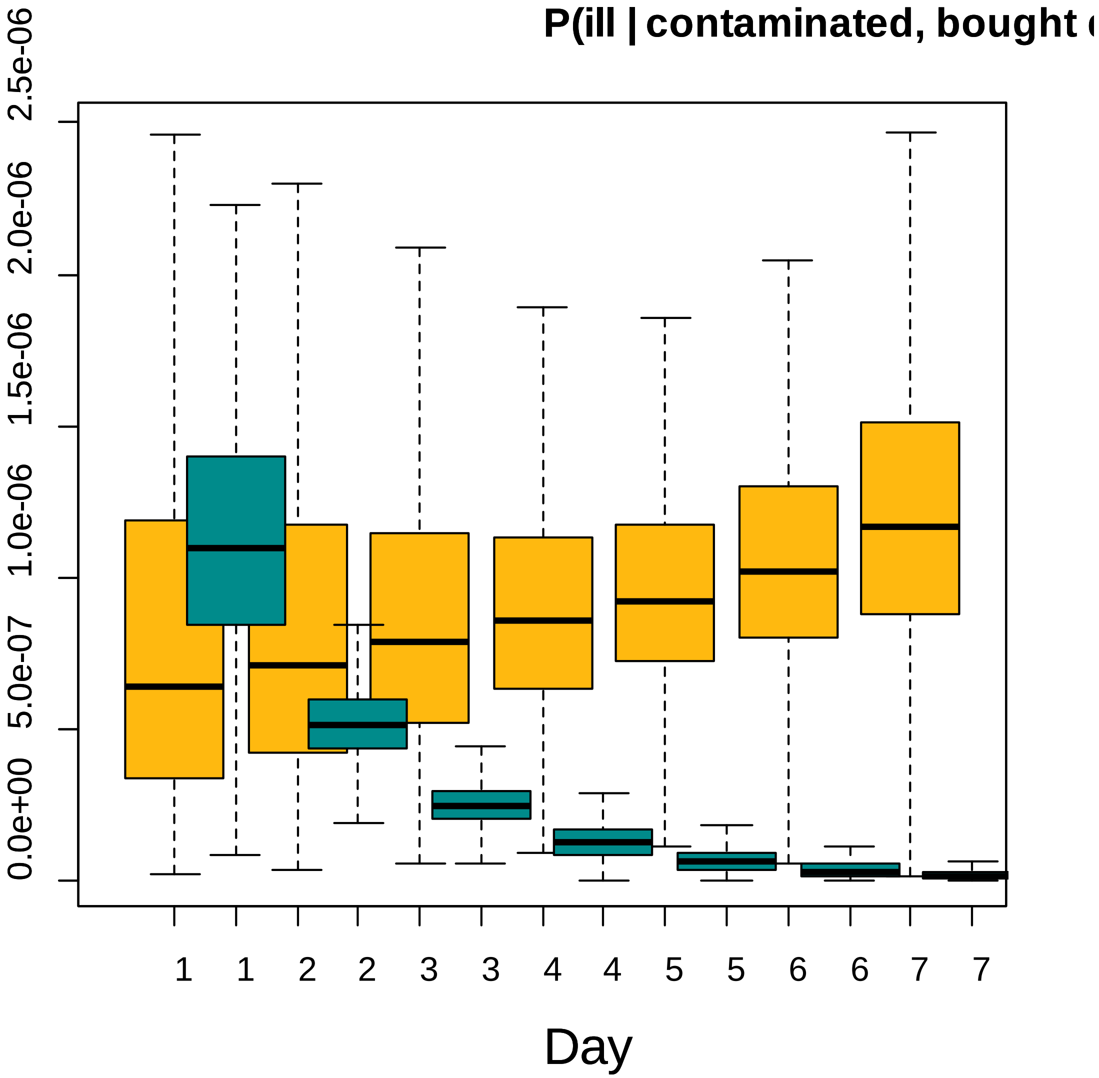
- Posterior daily probability of illness, if you DO eat food that was bought contaminated on day 1.

- Growth makes it increasingly risky.
  - But only if you eat it, and if you are still susceptible (i.e. not yet infected).



But you can stop eating it any day!

- Next: take into account the daily probability to continue consumption, and the survival probability to avoid infection (probability to be still susceptible for the first infection)

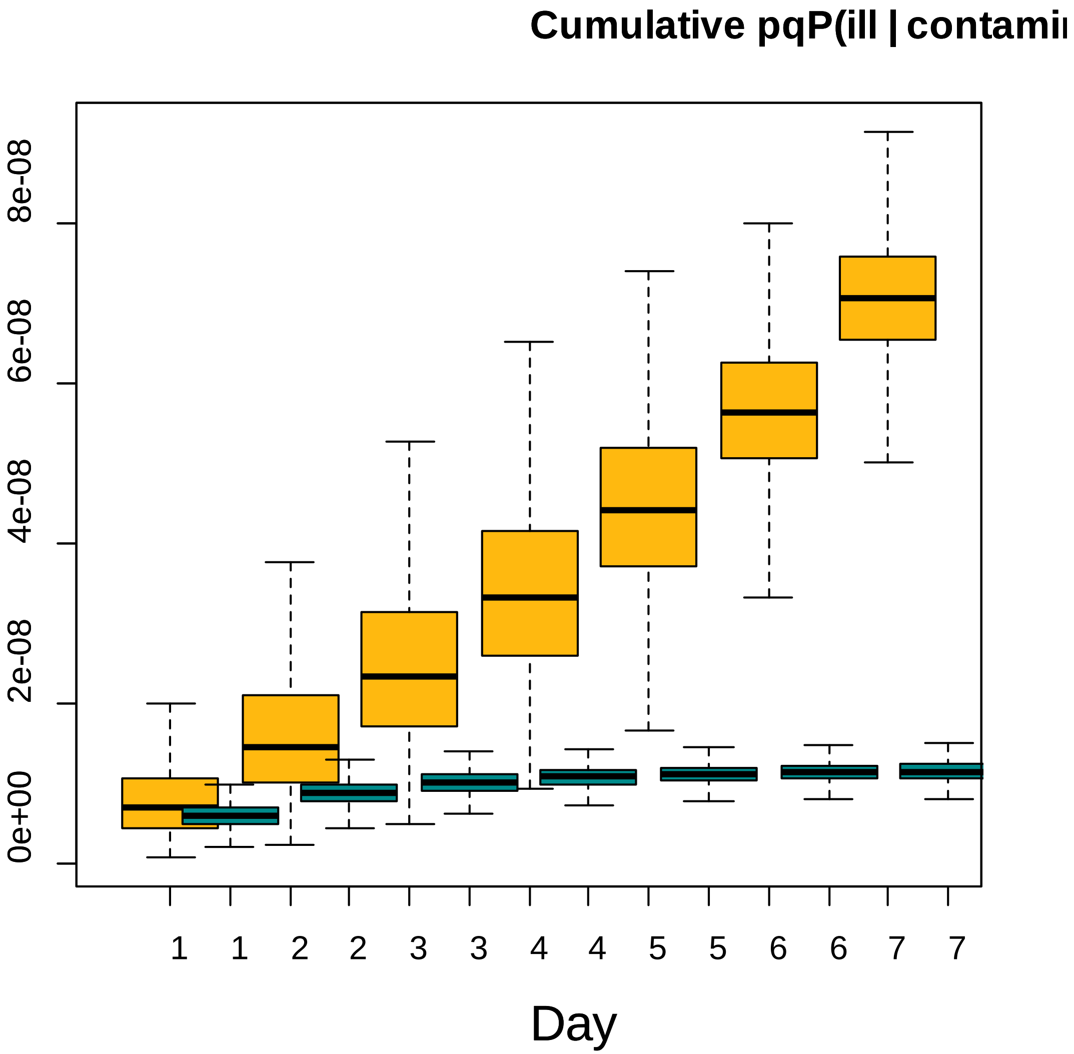


What is the cumulating effect?

- Cumulative probability of illness can reach a limit < pq < 1.

- A race of two opposite effects:
  - Bacteria growth *versus* quit eating a contaminated product.

- And the winner is?



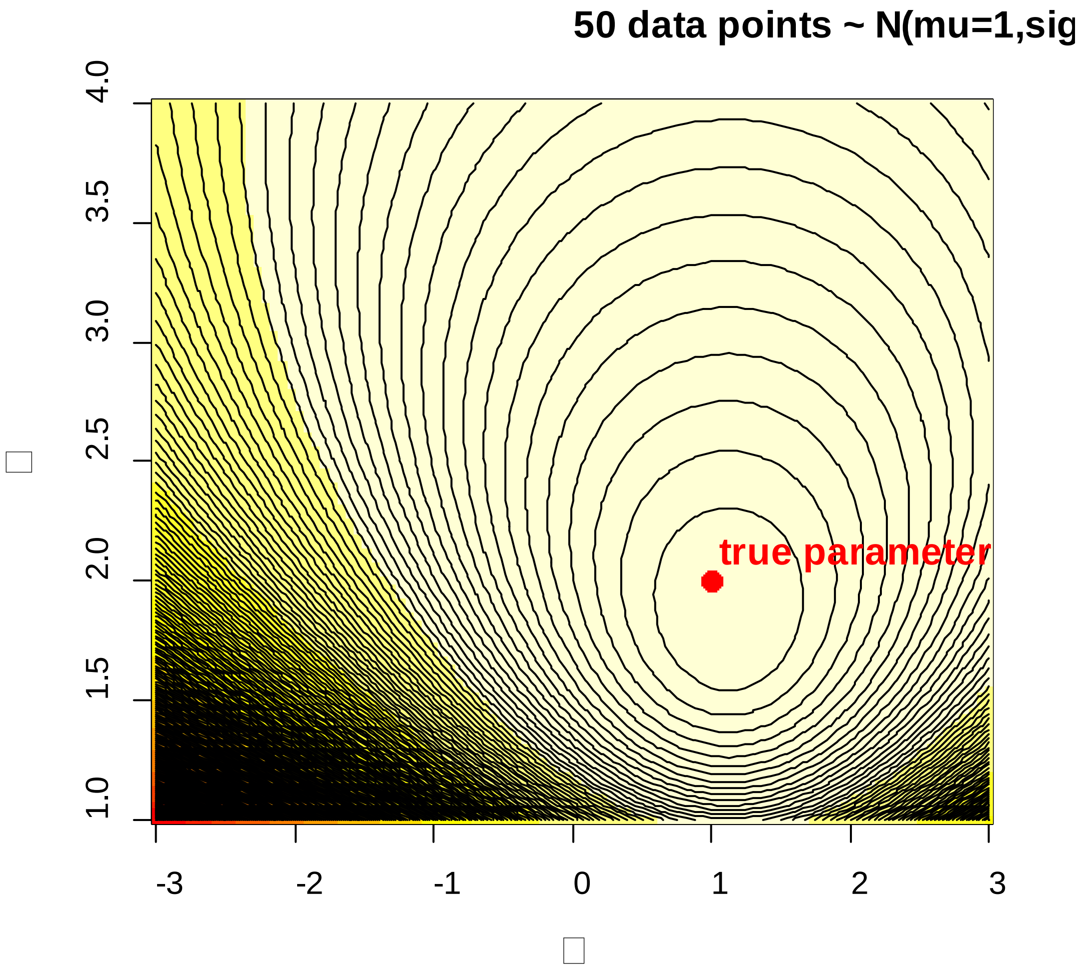
In this study we had both exact concentrations (CFU/g) and values below LOQ

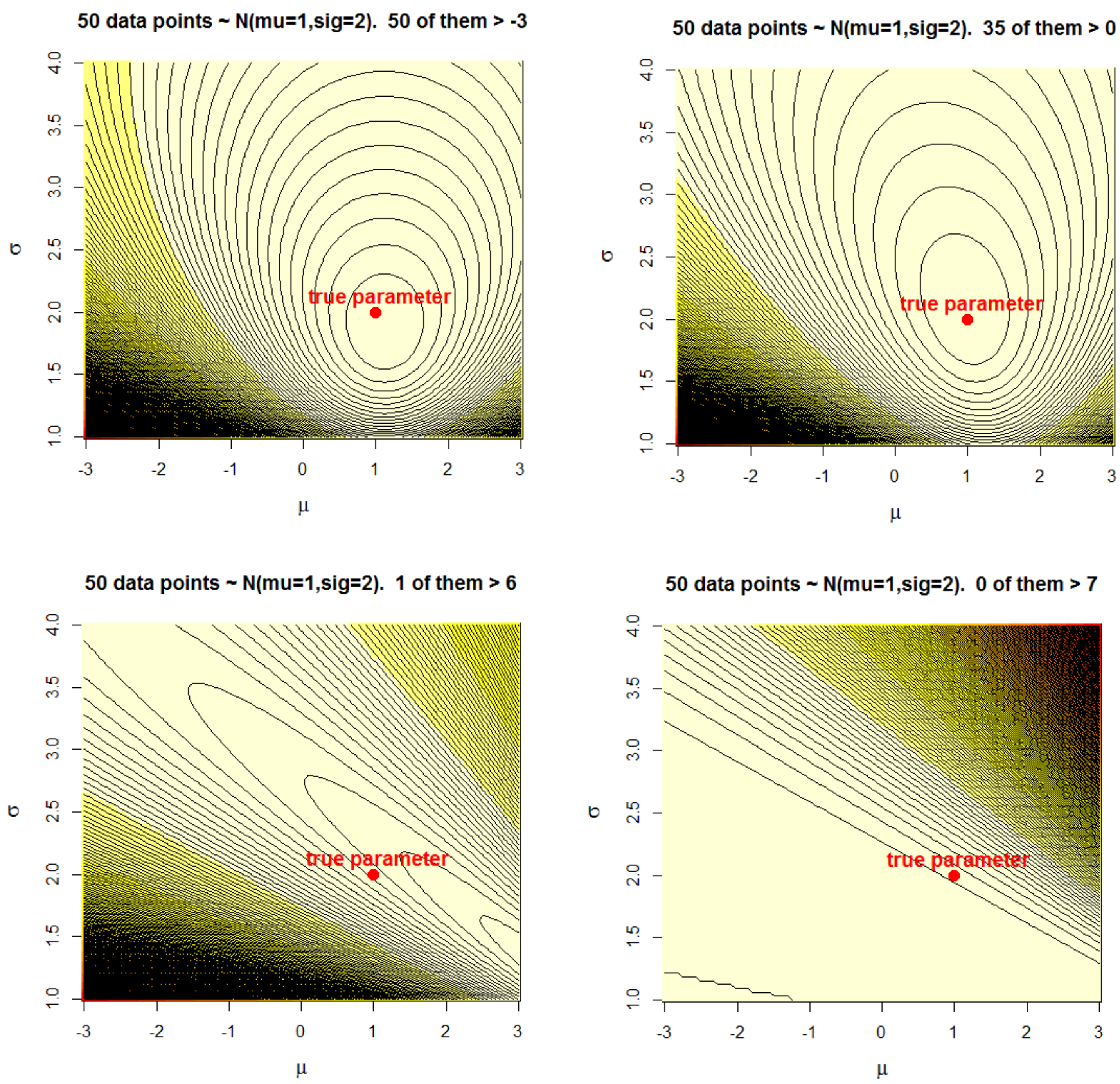
- Posterior distribution is based on full likelihood function.
  - $L(\mu,\sigma) = \prod P(y_i \mid \mu,\sigma^2\ ) \times \prod P(y_i < c \mid \mu,\sigma^2)$

- Example: likelihood function contours if no values below LOQ: →

 (same as posterior distribution contours if uniform prior)

- What if most, or all, measurements are below LOQ?
  - i.e. left censored: $y < c = LOQ$





50 data points ~ N(mu=1,sig=2). 50 of them > -3

50 data points ~ N(mu=1,sig=2). 35 of them > 0

50 data points ~ N(mu=1,sig=2). 7 of them > 3

50 data points ~ N(mu=1,sig=2). 1 of them > 6

50 data points ~ N(mu=1,sig=2). 0 of them > 7

This is what happens to parameter uncertainty when increasingly more data points fall below the censoring limit.

In applications, often very large proportion of concentration measurements are below LOQ.

DATA: CFU-values or original plate counts?

- CFU/g -value itself is an estimate from laboratory.
- Original data: plate counts from dilution series.
- Either way: Bayesian model accounting for true zeros and small concentrations that may lead to apparent zeros. (zero-inflated models).

- NOTE: we can only use the data we have.

→The model should reflect this.

So what then?

- *Bayesian methods have already been long used in many applications for food safety risk assessment, both chemical and microbiological.*

- *Not always called or known to be ”Bayesian”.*
  - *E.g. simulating outcomes X from model P(X|θ) with a range of parameter values θ (randomly from uniform distribution P(θ)), then selecting those parameters that led to a desired outcome X=x.*

  →this is the simplest Monte Carlo method for a posterior distribution P(θ|X=x), and the simplest ABC-method (without the ”A”).

- *Potential still not fully exploited.*
- *Increase probabilistic problems in basic training?*

Ref.

- *Ranta, Maijala: A probabilistic transmission model of Salmonella in the primary broiler production chain. Risk Analysis 2002, Vol 22, n1: 47-58.*
- *Smid, Swart, Havelaar, Pielaat: A practical framework for the construction of a biotracing model: application to Salmonella in the pork slaughter chain. Risk Analysis 2011, Vol 31, n9: 1434-1450*
- *Ranta, Lindqvist, Hansson, Tuominen, Nauta: A Bayesian approach to the evaluation of risk-based microbiological criteria for Campylobacter in broiler meat. The Annals of applied statistics 2015. Vol 9, n3: 1415-1432.*
- *Pella, Masuda: Bayesian methods for analysis of stock mixtures from genetic gharacters. Fish Bull 2001. 99: 151-167.*
- *Corander, Cui, Koski, Sirén: Have I seen you before? Principles of Bayesian predictive classification revisited. Statistics and computing 2013, Vol 23, issue 1: 59-73.*
- *Miller, Marshall, French, Jewell: SourceR: Classification and source attribution of infectious agents among heterogeneous populations. PLoS Comput Biol. 2017, 13 (5): e1005564.*
- *Pasonen, Ranta, Tapanainen, Valsta, Tuominen: Listeria monocytogenes risk assessment with a repeated exposure model. To be submitted.*
- *Pouillot, Hoelzer, Chen, Dennis: Listeria monocytogenes Dose Response revisited – incorporating adjustments for variability in strain virulence and host susceptibility. Risk Analysis 2015, vol 35, n1: 90-108.*
- *Pouillot, Hoelzer, Chen, Dennis: Estimating probability distributions of bacterial concentrations in food based on data generated using the most probable number (MPN) method for use in risk assessment. Food Control 2013, 29: 350-357.*

Thank you !

Tack !

Kiitos !