


# Optimized sample sizes for analyzing the genetic heterogeneity of *Escherichia coli* isolated from sheep and calves

Dörte Döpfer<sup>1</sup>, R.N. Zadoks<sup>2</sup>, W. Buist<sup>1</sup>, and B. Engel<sup>1</sup>



(1) Quantitative Veterinary Epidemiology Group, Animal Sciences Group of Wageningen UR, The Netherlands  
 (2) Department of Food Science and Quality Milk Control Services, Cornell University, Ithaca/NY, USA




7<sup>th</sup> of June, 2007

- Introduction
- genetic heterogeneity in multiple isolates of one sample is a common problem in microbiology
- the number of isolates tested, genotyped, virulotyped etc. is often determined by:
  - convenience
  - habit
  - resources


there is a need to optimise the sample sizes in order to make optimal usage of available resources.


- the problem
- of genetic heterogeneity in multiple isolates of a sample





- need to answer:
  - How many isolates have to be genotyped in order to be 95% confident that all genotypes are found?



- the statistical problem:
  - of genetic heterogeneity in multiple isolates of one sample



- a statistical "occupancy problem":
  - how many ways are there to distribute 5 jackets over 10 hangers in a cloakroom?
  - (Johnson & Kotz 1969)
- need for the probability of finding all genotypes
  - while analyzing a fraction of isolates from one sample





- a Bayesian approach:
  - calculate the probability for finding *j* genotypes given that truly *i* present in *N* isolates genotyped (Johnson & Kotz 1969)

$$P_{N(j|i)} = \binom{i}{j} \sum_{r=0}^j (-1)^r \binom{j}{r} \left(\frac{j-r}{i}\right)^N$$

need for a probability matrix

- use real-world data
  - see next slides
  - need sheep
- built priors consulting an expert...
  - "prior" knowledge:
    - expert expects an average of 6 genotypes in 10 isolates,
    - (theta 1.... theta k) ~Dirichlet(1/k....1/k) ← prior distribution
    - Pre-assumption: all isolates are equally likely to be found!
    - Altekruse et al. 2003, Singer et al. 2000




- the probability matrix:
  - using Genstat, SAS, Excel....per sample: f.e. N=5, i=5, j=4

$$P_{N(j|i)} = \binom{i}{j} \sum_{r=0}^j (-1)^r \binom{j}{r} \left(\frac{j-r}{i}\right)^N$$

1.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000
0.500000	0.500000	0.000000	0.000000	0.000000
0.333333	0.666667	0.000000	0.000000	0.000000
0.250000	0.750000	0.000000	0.000000	0.000000
0.200000	0.800000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000
0.250000	0.750000	0.000000	0.000000	0.000000
0.111111	0.666667	0.222222	0.000000	0.000000
0.062500	0.382500	0.375000	0.000000	0.000000
0.040000	0.480000	0.480000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000
0.125000	0.875000	0.000000	0.000000	0.000000
0.037037	0.518519	0.444444	0.000000	0.000000
0.016250	0.328125	0.562500	0.093750	0.000000
0.008000	0.224000	0.576000	0.192000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000
0.062500	0.937500	0.000000	0.000000	0.000000
0.012348	0.370370	0.617284	0.000000	0.000000
0.003906	0.172781	0.555556	0.243750	0.000000
0.001600	0.096000	0.480000	0.384000	0.000000

N: total # of isolates typed  
 j: # of types observed  
 i: # of types truly present

- the occupancy problem – cloakroom statistics



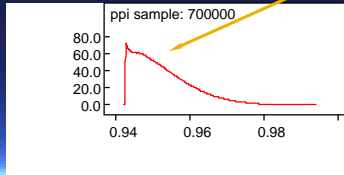
■ assemble.....

- expert opinion as prior knowledge,
- data and
- probabilities

- to calculate the **posterior likelihood distribution** for finding all genotypes in a given sample while analyzing a fraction N

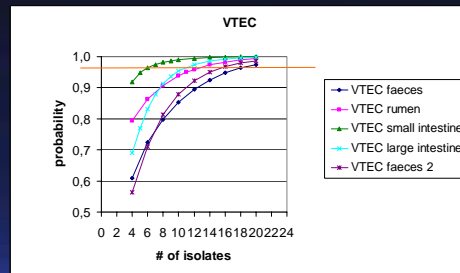
$$p(N) = \sum_{i=1}^k \theta_i P_N(i | i)$$

- using a Gibbs sampler (Spiegelhalter et al. 2000)



■ the results I:

- calculating sample sizes for genotyping multiple isolates of one faecal sample in different segments of the ovine GI tract

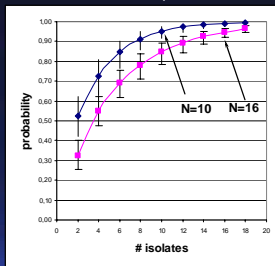


results:  
 faeces 1: 16 isolates  
 faeces 2: 14  
 rumen: 11  
 large int: 10  
 small int: 5

- probability of finding all strains of *E. coli* when genotyping N isolates per sample (incl. 95% credible intervals) for different segments of the ovine GI tract

■ the results II:

- calculating sample sizes for virulotyping multiple isolates of one faecal sample based on 100 samples of 10 colonies each from Geue et al. 2002 (faecal non type-specific *E. coli* from calves)



results:  
 faeces 1: 16 isolates  
 (vt1, vt2, eae, hly, katP, espP, colicin)  
 faeces 2: 10 isolates  
 (vt1, vt2, eae, hly, cif, efa1, saa)

- probability of finding all strains of *E. coli* when genotyping N isolates per sample (incl. 95% credible intervals) for faecal *E. coli* from slaughter calves

■ Conclusions

- general problem in microbiology
- applicable to surveys with the purpose of genotyping, virulotyping etc.
- use stepwise iterations to update sample sizes in surveys –
- use less expensive typing methods first and more sophisticated ones later (microarrays, MLST...)
- optimise the allocation of resources using expert opinion, real-world data and probability methods

■ Thank you for your attention!!!

dorte.dopfer@wur.nl