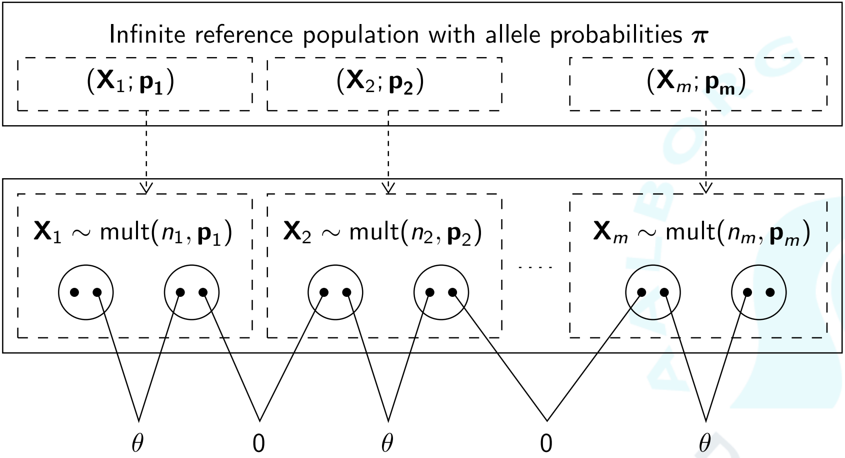
Overdispersion in allelic counts and θ-correction in forensic genetics

Torben Tvedebrink
PhD Student, MSc

Department of Mathematical Sciences
Aalborg University - Denmark

Oral presentation: O 30
September 18 2009 - ISFG2009 Congress - Buenos Aires

Model setup



$$\mathbf{p}_i \sim \text{Dirichlet}(\boldsymbol{\pi}, \theta), \quad i = 1, \dots, m$$

Dirichlet-multinomial distribution

- The unconditional distribution of $\mathbf{X}_i$ is a Dirichlet-multinomial distribution with parameters $\boldsymbol{\pi}$ and θ:

$$P(\mathbf{X}_i = \mathbf{x}_i) = \binom{n_i}{\mathbf{x}_i} \frac{\Gamma([1-\theta]/\theta)}{\Gamma(n_i + [1-\theta]/\theta)} \prod_{j=1}^{k} \frac{\Gamma\left(x_{ij} + \pi_j[1-\theta]/\theta\right)}{\Gamma\left(\pi_j[1-\theta]/\theta\right)},$$

- The mean and variance of a Dirichlet-multinomial variable is:

$$\mathbb{E}(\mathbf{X}_i) = n_i\boldsymbol{\pi}$$
$$\mathbb{V}(\mathbf{X}_i) = n_i(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top)[1 + \theta(n_i - 1)],$$

- I.e. a multinomial variance with an additional variance term governed by the overdispersion parameter θ.

Dependent alleles

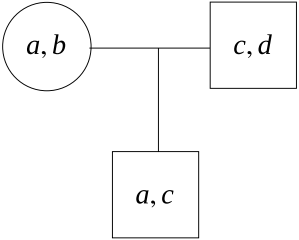
From the model assumptions and the Dirichlet-multinomial distribution we may recover David Balding's sampling formula:

$$P(x_j + 1 | x_j, n) = \frac{x_j\theta + (1-\theta)\pi_j}{1 + (n-1)\theta},$$

i.e. the probability of observing a future j allele only depend on the total number of sampled alleles, n, and the number of j alleles among those, x_j.

Paternity Index

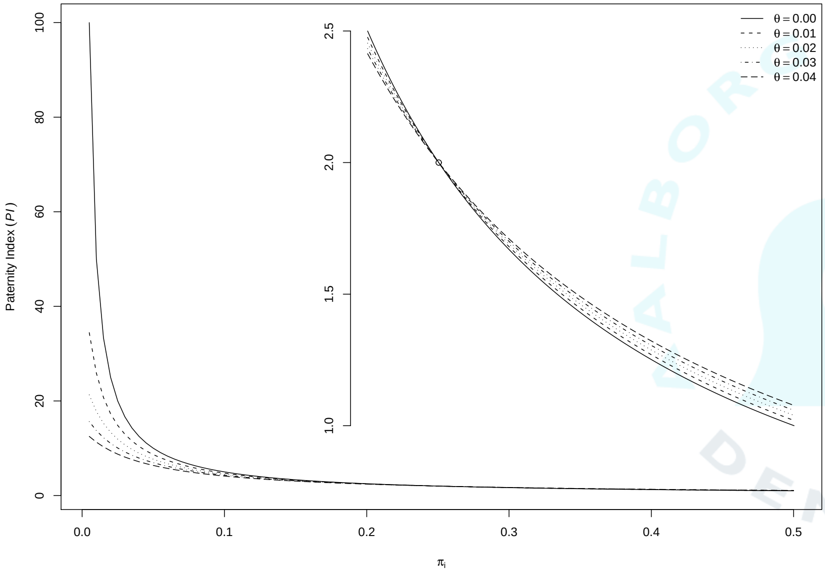
Consider the following trio:



Using the formula from the previous slide the paternity index is:

$$PI(\theta) = \frac{1 + (n-1)\theta}{2[x_c\theta + (1-\theta)\pi_c]} = \frac{1 + 3\theta}{2[\theta + (1-\theta)\pi_c]},$$

Paternity Index



Maximum likelihood estimation

- ML estimation under linear constraints: $\sum_{j=1}^{k} \pi_j = 1$.
- Reparameterisation:

$$\gamma_j = \pi_j \frac{1-\theta}{\theta} \quad \text{and} \quad \gamma_+ = \sum_{j=1}^{k} \gamma_j = \frac{1-\theta}{\theta}$$

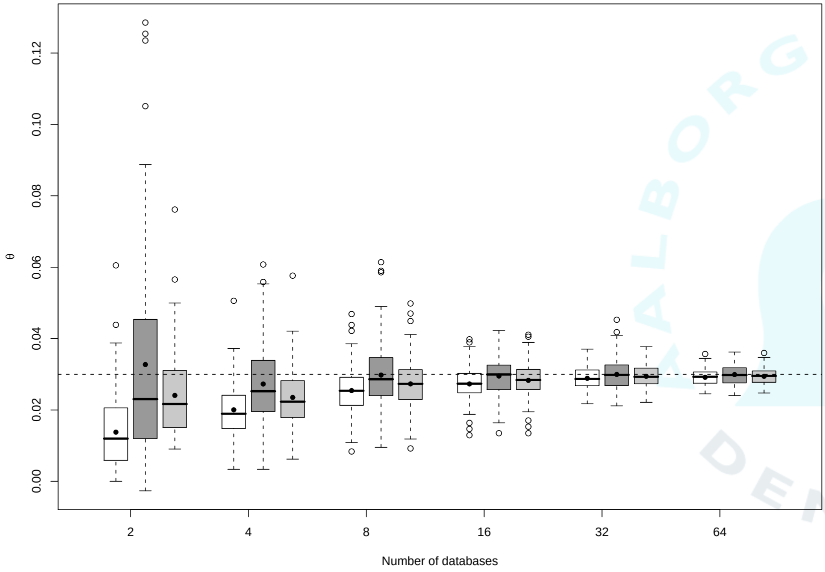
- Log-likelihood:

$$\ell(\boldsymbol{\gamma}; \mathbf{x}) = \sum_{i=1}^{m} \sum_{j=1}^{k} \sum_{r=1}^{x_{ij}} \log\{\gamma_j + r - 1\} - \sum_{i=1}^{m} \sum_{r=1}^{n_i} \log\{\gamma_+ + r - 1\},$$

Simulating from Dirichlet-multinomial

- Simulate from a known distribution to validate implementation.
- 100 simulations with $\theta = 0.03$ and known allele probabilities.
- Compare MLE with the Method of Moment (MoM) estimator of Weir and Hill (2002)
  (Previously derived in Weir and Cockerham, 1984).

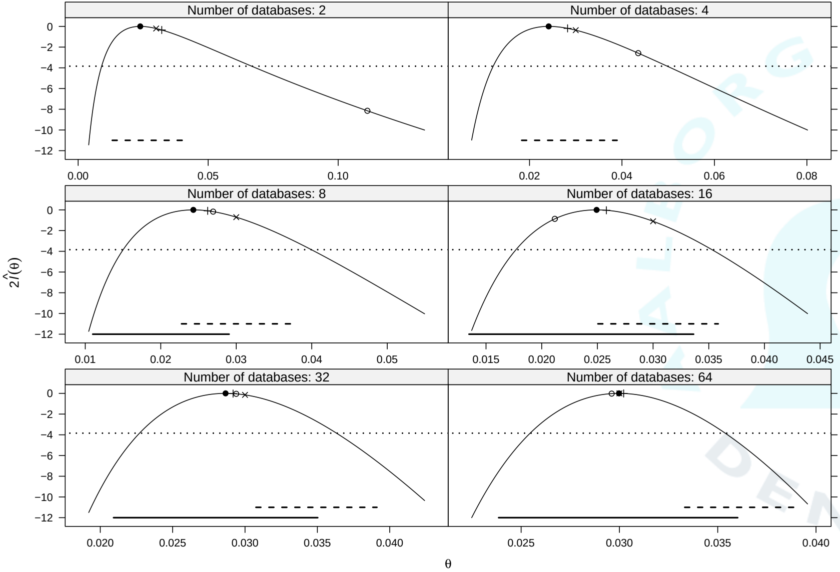
Simulations



Profile log-likelihood

- Log-likelihood used for evaluating: $\hat{\ell}(\theta_0; \mathbf{x}) = \max_{\boldsymbol{\pi}} \ell(\boldsymbol{\pi}, \theta_0; \mathbf{x})$
- Lagrange multiplier λ: $\tilde{\ell}(\boldsymbol{\pi}, \theta; \mathbf{x}) = \ell(\boldsymbol{\pi}, \theta; \mathbf{x}) + \lambda(\theta_0 - \theta)$
- Plotting the $\hat{\ell}(\theta_0; \mathbf{x})$ for $\theta_0 \in [\hat{\theta} - a \,;\, \hat{\theta} + b]$ can be used to investigate bias of MLE.
- The profile log-likelihood may also be used to construct approximative confidence intervals for θ.

Simulated data



Hypothesis testing

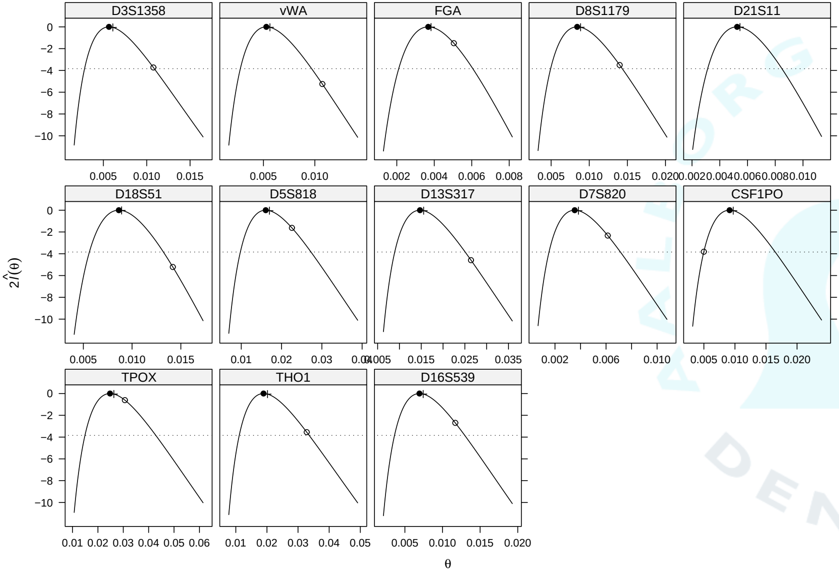
The MLE implementation may be used to test hypothesis:

- Test the hypothesis: Equal θ-values for multiple loci
- Test the hypothesis: $\theta = 0$

FBI data

- Data from six US sub-populations published by Budowle and Moretti, 1999.
- The data contain CODIS core DNA profiles from approximately 1100 individuals distributed on African American, Bahamian, Jamaican, Trinidad, Caucasian and Hispanic sub-populations.

FBI data - Profile log-likelihoods



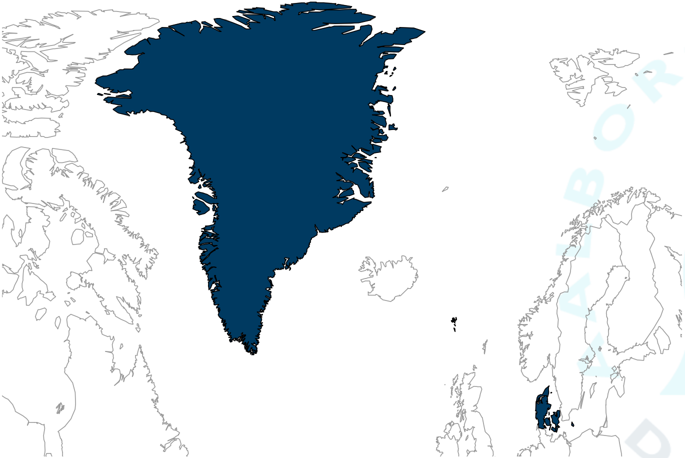
FBI data - Estimates

Locus	$\tilde{\theta}_{\text{MoM}}$	$SE(\tilde{\theta})$	$\hat{\theta}_{\text{MLE}}$	$SE(\hat{\theta})$	95%-CI for $\hat{\theta}$	Post.mean
D3	0.0108	0.0085	0.0056	0.0020	(0.0028 ; 0.0110)	0.0061
vWA	0.0107	0.0085	0.0053	0.0017	(0.0027 ; 0.0098)	0.0056
FGA	0.0050	0.0051	0.0037	0.0010	(0.0021 ; 0.0061)	0.0038
D8	0.0140	0.0106	0.0084	0.0024	(0.0049 ; 0.0145)	0.0089
D21	0.0126	0.0097	0.0053	0.0013	(0.0031 ; 0.0086)	0.0055
D18	0.0142	0.0107	0.0086	0.0019	(0.0056 ; 0.0133)	0.0089
D5	0.0226	0.0157	0.0161	0.0042	(0.0097 ; 0.0276)	0.0170
D13	0.0264	0.0180	0.0147	0.0040	(0.0088 ; 0.0254)	0.0156
D7	0.0061	0.0056	0.0035	0.0013	(0.0015 ; 0.0072)	0.0038
CSF	0.0050	0.0049	0.0091	0.0026	(0.0049 ; 0.0167)	0.0097
TPOX	0.0306	0.0205	0.0248	0.0066	(0.0147 ; 0.0433)	0.0263
TH01	0.0328	0.0217	0.0189	0.0054	(0.0110 ; 0.0340)	0.0202
D16	0.0117	0.0091	0.0069	0.0023	(0.0036 ; 0.0131)	0.0074

FBI data - Hypothesis testing

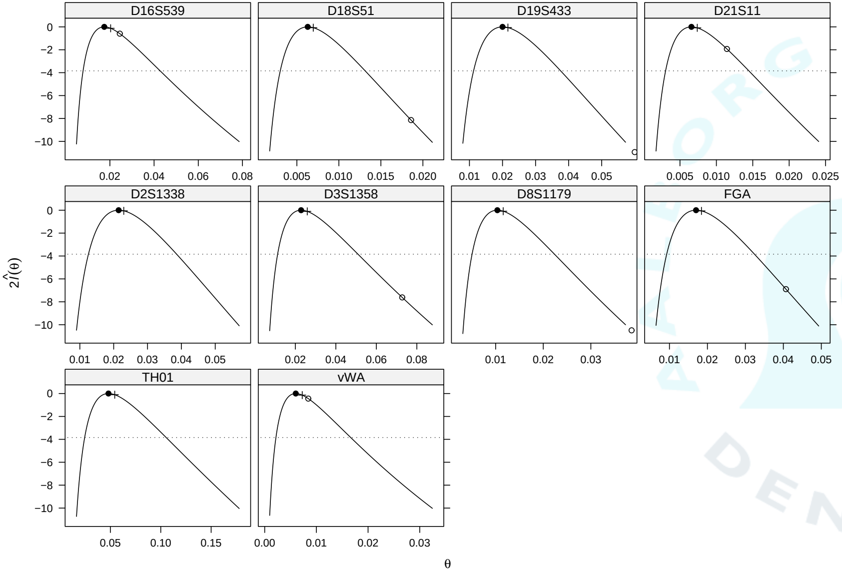
Locus	$\tilde{\theta}_{\text{MoM}}$	$SE(\tilde{\theta})$	$\hat{\boldsymbol{\theta}}_{\textbf{MLE}}$	$SE(\hat{\theta})$	95%-CI for $\hat{\theta}$	Post.mean
D3	0.0108	0.0085	**0.0056**	0.0020	(0.0028 ; 0.0110)	0.0061
vWA	0.0107	0.0085	**0.0053**	0.0017	(0.0027 ; 0.0098)	0.0056
FGA	0.0050	0.0051	**0.0037**	0.0010	(0.0021 ; 0.0061)	0.0038
D8	0.0140	0.0106	**0.0084**	0.0024	(0.0049 ; 0.0145)	0.0089
D21	0.0126	0.0097	**0.0053**	0.0013	(0.0031 ; 0.0086)	0.0055
D18	0.0142	0.0107	**0.0086**	0.0019	(0.0056 ; 0.0133)	0.0089
D5	0.0226	0.0157	**0.0161**	0.0042	(0.0097 0.0276)	0.0170
D13	0.0264	0.0180	**0.0147**	0.0040	(0.0088 0.0254)	0.0156
D7	0.0061	0.0056	**0.0035**	0.0013	(0.0015 ; 0.0072)	0.0038
CSF	0.0050	0.0049	**0.0091**	0.0026	(0.0049 ; 0.0167)	0.0097
TPOX	0.0306	0.0205	**0.0248**	0.0066	(0.0147 0.0433)	0.0263
TH01	0.0328	0.0217	**0.0189**	0.0054	(0.0110 0.0340)	0.0202
D16	0.0117	0.0091	**0.0069**	0.0023	(0.0036 ; 0.0131)	0.0074

Danish data



Data made available by The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark.

Danish data - Profile log-likelihoods



Danish data - Estimates

Locus	$\tilde{\theta}_{\text{MoM}}$	$SE(\tilde{\theta})$	$\hat{\theta}_{\text{MLE}}$	$SE(\hat{\theta})$	95%-CI for $\hat{\theta}$	Post.mean
D3	0.0728	0.0700	0.0228	0.0093	(0.0114 ; 0.0507)	0.0259
vWA	0.0084	0.0112	0.0060	0.0030	(0.0023 ; 0.0166)	0.0073
D16	0.0245	0.0266	0.0175	0.0077	(0.0080 ; 0.0426)	0.0203
D2	0.0994	0.0919	0.0215	0.0067	(0.0129 ; 0.0382)	0.0230
D8	0.0386	0.0397	0.0104	0.0042	(0.0052 ; 0.0222)	0.0116
D21	0.0114	0.0141	0.0066	0.0025	(0.0030 ; 0.0145)	0.0074
D18	0.0186	0.0210	0.0063	0.0024	(0.0031 ; 0.0131)	0.0069
D19	0.0600	0.0589	0.0200	0.0064	(0.0116 ; 0.0372)	0.0216
TH01	0.2059	0.1653	0.0481	0.0195	(0.0250 ; 0.1048)	0.0543
FGA	0.0407	0.0417	0.0170	0.0052	(0.0095 ; 0.0323)	0.0184

Danish data - Hypothesis testing

Locus	$\tilde{\theta}_{\text{MoM}}$	$SE(\tilde{\theta})$	$\hat{\boldsymbol{\theta}}_{\textbf{MLE}}$	$SE(\hat{\theta})$	95%-CI for $\hat{\theta}$	Post.mean
D3	0.0728	0.0700	**0.0228**	0.0093	(0.0114 0.0507)	0.0259
vWA	0.0084	0.0112	**0.0060**	0.0030	(0.0023 ; 0.0166)	0.0073
D16	0.0245	0.0266	**0.0175**	0.0077	(0.0080 0.0426)	0.0203
D2	0.0994	0.0919	**0.0215**	0.0067	(0.0129 0.0382)	0.0230
D8	0.0386	0.0397	**0.0104**	0.0042	(0.0052 0.0222)	0.0116
D21	0.0114	0.0141	**0.0066**	0.0025	(0.0030 ; 0.0145)	0.0074
D18	0.0186	0.0210	**0.0063**	0.0024	(0.0031 ; 0.0131)	0.0069
D19	0.0600	0.0589	**0.0200**	0.0064	(0.0116 0.0372)	0.0216
TH01	0.2059	0.1653	**0.0481**	0.0195	(0.0250 0.1048)	0.0543
FGA	0.0407	0.0417	**0.0170**	0.0052	(0.0095 0.0323)	0.0184

R-package: `dirmult`

A R-package, `dirmult`, with functions for:

- Estimating in a Dirichlet-multinomial distribution
- Simulating from a Dirichlet-multinomial distribution
- Compute profile log-likelihood functions
- Hypothesis testing for equality of θ across loci or testing $\theta = 0$

is available online at `www.cran.r-project.org`.

References

B. Budowle and T. R. Moretti (1999). Genotype Profiles for Six Population Groups at the 13 CODIS Short Tandem Repeat Core Loci and Other PCR-Based Loci. *Forensic Science Communications*

Available online:
`http://www.fbi.gov/hq/lab/fsc/backissu/july1999/budowle.htm`

B. S. Weir, and W. G. Hill (2002). Estimating F-statistics. *Annual Review of Genetics* 36, 721-750.

B. S. Weir and C. C. Cockerham (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358-1370.