



Assessing seeds with stacks in conventional seed lots

ISTA GMO TF – ISTA Statistics Committee

Jean-Louis Laffont, Sylvain Grégoire, Kirk Remund



Acknowledgements

- **Christoph Haldemann - ISTA GMO TF**
- **Colleagues at Pioneer and Monsanto.**

In particular:

**Bonnie Hong and Kevin Wright (Pioneer)
Tim Perez (Monsanto)**



The problem

Using an appropriate testing plan for semi-quantitative methods, we estimate the % of GM seeds in the lot to be equal to 0.4%

Seed lot (40 tons)



From these 0.4% GM seeds, which percentage have

- a single trait?
- two traits?
- three traits?
- ...





The problem – Double stack assessment

Seed lot: true total

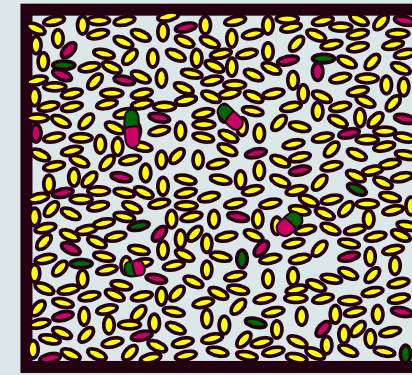
AP proportion = θ

θ_0 : conventional seeds

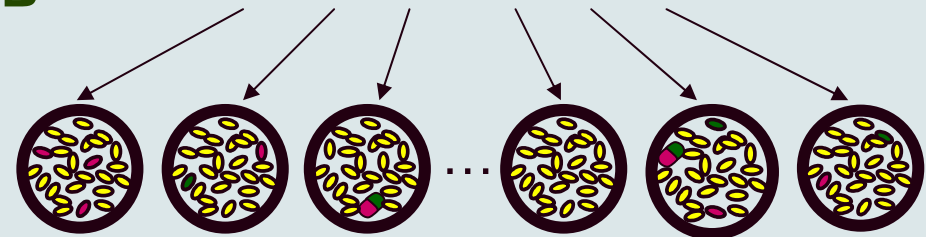
θ_A : seeds with event A

θ_B : seeds with event B

θ_{AB} : seeds with events A & B



n pools of
 m seeds



Qualitative test for A
and B on each pool

Event	Pool 1	Pool 2	Pool 3	...	Pool (m-2)	Pool (m-1)	Pool m
A	+	+	+	...	-	+	+
B	-	+	+	...	-	+	+

Observed pattern: n_0 negative pools for A and B, n_A positive pools for A only, n_B positive pools for B only, n_{AB} positive pools for both A and B

Can we estimate θ_A , θ_B and θ_{AB}
from this pattern?



Solution



Difficulty: a pool can be positive for **A** and **B** either because there are seeds with **A** only and seeds with **B** only **or** only seeds with **A** and **B**

Example: Pools of 300 seeds

$$\theta_A = 0.3\% , \theta_B = 0.3\% , \\ \theta_{AB} = 0\%$$



Prob the pool is positive for **A** and **B** = **35.2%**

$$\theta_A = 0\% , \theta_B = 0\% , \\ \theta_{AB} = 0.3\%$$



Prob the pool is positive for **A** and **B** = **59.4%**

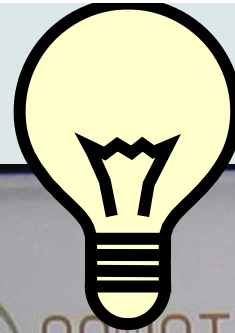


Solution





Solution



Maximum-likelihood estimation:

Find the parameter values $\hat{\theta}_A$, $\hat{\theta}_B$, $\hat{\theta}_{AB}$ that make the observed pattern (n_A, n_B, n_{AB}) most likely.



Maximum-likelihood estimation: mathematically

Likelihood function for double stack assessment:

When observing the pattern (n_0, n_{AB}, n_A, n_B) , the likelihood function can be written as:

$$L(\theta_A, \theta_B, \theta_{AB}) = \frac{n!}{n_0! n_A! n_B! n_{AB}!} (1 - p_A - p_B - p_{AB})^{n_0} p_A^{n_A} p_B^{n_B} p_{AB}^{n_{AB}}$$

$$\propto (1 - p_A - p_B - p_{AB})^{n_0} p_A^{n_A} p_B^{n_B} p_{AB}^{n_{AB}}$$

Multinomial distribution

where:

$$p_0 = \text{p(the pool has only conventional seeds)} = \theta_0^m$$

$$p_A = \text{p(the pool is positive for A, negative for B)} = [\theta_A + \theta_0]^m - p_0$$

$$p_B = \text{p(the pool is positive for B, negative for A)} = [\theta_B + \theta_0]^m - p_0$$

$$p_{AB} = \text{p(the pool is positive for both A and B)} = 1 - p_0 - p_A - p_B$$

Log-likelihood is:

$$l(\theta_A, \theta_B, \theta_{AB}) = n_0 \ln(1 - p_A - p_B - p_{AB}) + n_A \ln(p_A) + n_B \ln(p_B) + n_{AB} \ln(p_{AB})$$

Finding the parameters $\theta_A, \theta_B, \theta_{AB}$ is done by maximizing the above log-likelihood and θ_0 is found by subtraction.

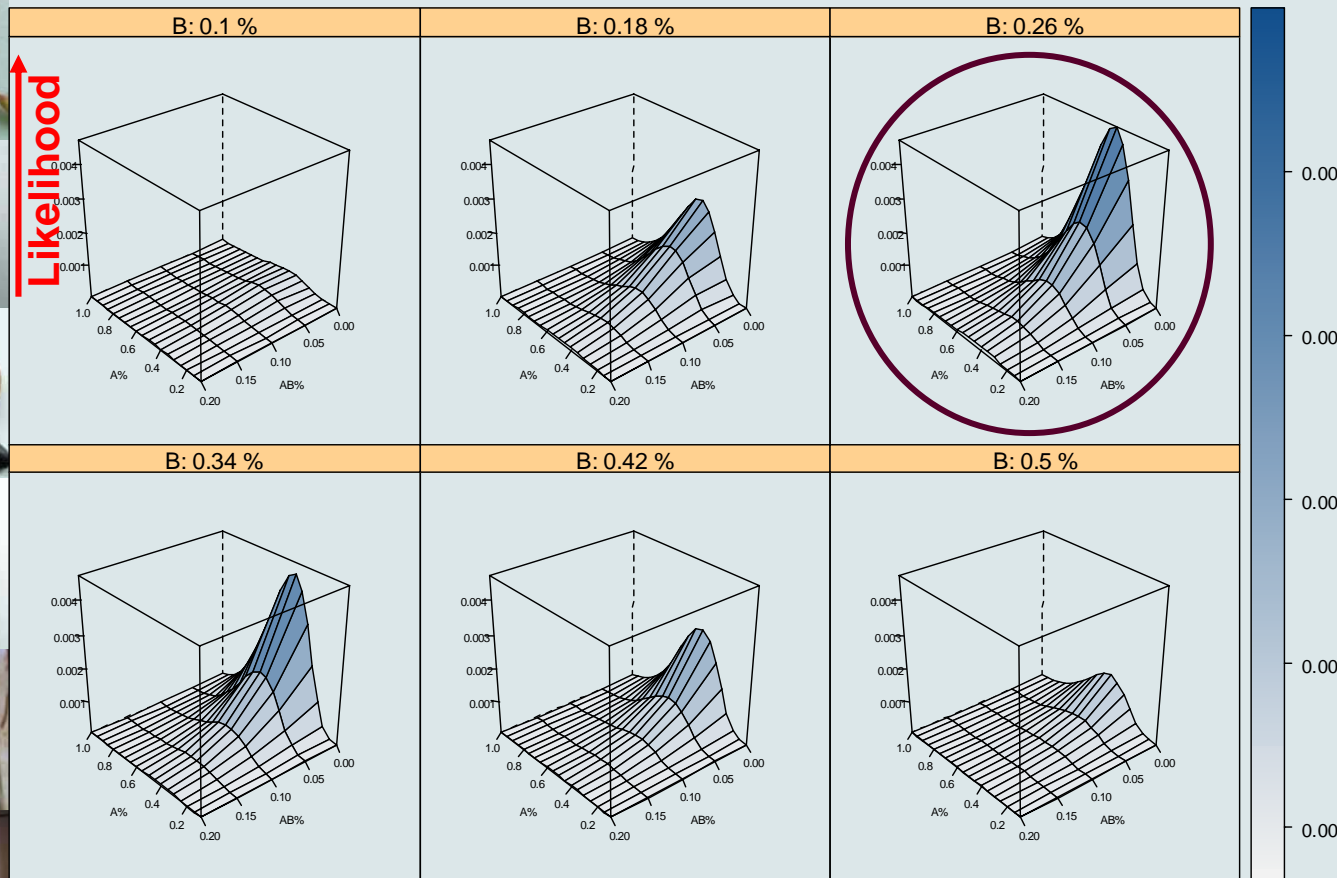
Many optimization methods are available in most statistical software to minimize the negative of the above log-likelihood (for example, NLP subroutines from SAS IML procedure, `optim()` function in R, `nlminb()` function in S-PLUS).



Maximum-likelihood estimation: illustration

20 pools of
150 seeds

Observed pattern: 5 negative pools for **A** and **B**
8 positive pools for **A** only
5 positive pools for **B** only
2 positive pools for both **A** and **B**



Maximum
Likelihood
for :

$$\hat{\theta}_A = 0.46\%$$

$$\hat{\theta}_B = 0.29\%$$

$$\hat{\theta}_{AB} = 0\%$$



Confidence Intervals for the proportions

→ Use of the property that maximum likelihood estimators asymptotically follow normal distribution to derive confidence intervals

In the case of double stack assessment, the approximate variance-covariance matrix \mathbf{V} of MLE vector $\hat{\theta} = (\hat{\theta}_A, \hat{\theta}_B, \hat{\theta}_{AB})'$ is the negative of the inverse of Hessian matrix of $l(\theta; y)$ evaluated at the point estimators,

$$V(\hat{\theta}) \approx [-H(\hat{\theta}; y)]^{-1} = \begin{pmatrix} \frac{\partial^2 l(y; \theta_A, \theta_B, \theta_{AB})}{\partial \theta_A^2} & \frac{\partial^2 l(y; \theta_A, \theta_B, \theta_{AB})}{\partial \theta_A \theta_B} & \frac{\partial^2 l(y; \theta_A, \theta_B, \theta_{AB})}{\partial \theta_A \theta_{AB}} \\ \frac{\partial^2 l(y; \theta_A, \theta_B, \theta_{AB})}{\partial \theta_A \theta_B} & \frac{\partial^2 l(y; \theta_A, \theta_B, \theta_{AB})}{\partial \theta_B^2} & \frac{\partial^2 l(y; \theta_A, \theta_B, \theta_{AB})}{\partial \theta_B \theta_{AB}} \\ \frac{\partial^2 l(y; \theta_A, \theta_B, \theta_{AB})}{\partial \theta_A \theta_{AB}} & \frac{\partial^2 l(y; \theta_A, \theta_B, \theta_{AB})}{\partial \theta_B \theta_{AB}} & \frac{\partial^2 l(y; \theta_A, \theta_B, \theta_{AB})}{\partial \theta_{AB}^2} \end{pmatrix}^{-1}_{\hat{\theta}=(\hat{\theta}_A, \hat{\theta}_B, \hat{\theta}_{AB})}$$

and standard deviations are squared root of the diagonal elements of \mathbf{V} . The $(100-\alpha)\%$ confidence intervals for $\theta_A, \theta_B, \theta_{AB}$ are:

$$LB(\theta_A) = \hat{\theta}_A - z_{1-\alpha/2} \text{diag}(\mathbf{V})_1 \quad UB(\theta_A) = \hat{\theta}_A + z_{1-\alpha/2} \text{diag}(\mathbf{V})_1$$

$$LB(\theta_B) = \hat{\theta}_B - z_{1-\alpha/2} \text{diag}(\mathbf{V})_2 \quad UB(\theta_B) = \hat{\theta}_B + z_{1-\alpha/2} \text{diag}(\mathbf{V})_2$$

$$LB(\theta_{AB}) = \hat{\theta}_{AB} - z_{1-\alpha/2} \text{diag}(\mathbf{V})_3 \quad UB(\theta_{AB}) = \hat{\theta}_{AB} + z_{1-\alpha/2} \text{diag}(\mathbf{V})_3$$

respectively.



Generalization to K -stack assessment

... possible!!!



...

$$p^{(0)} = \text{p(the pool has only conventional seeds)} = \theta^{(0)m}$$

$$p_i^{(1)} = \left[\theta^{(i)} + \theta^{(0)} \right]^m - p^{(0)}$$

$$p_{i_1, i_2}^{(2)} = \text{p(the pool is positive for both trait } i_1 \text{ and trait } i_2, \text{ negative for other traits)}$$

$$= \left[\sum \theta_j^{(1)} + \theta_{i_1, i_2}^{(2)} + \theta^{(0)} \right]^m - p^{(0)} - \sum p_j^{(1)} \text{ where } j = i_1, i_2$$

$$p_{i_1, i_2, i_3}^{(3)} = \text{p(the pool is positive for trait } i_1 \text{ and trait } i_2 \text{ and trait } i_3, \text{ negative for other traits)}$$

$$= \left[\sum \theta_j^{(1)} + \sum \theta_{j_1, j_2}^{(2)} + \theta_{i_1, i_2, i_3}^{(3)} + \theta^{(0)} \right]^m - p^{(0)} - \sum p_j^{(1)} - \sum p_{j_1, j_2}^{(2)}$$

where $j = i_1, i_2, i_3, (j_1, j_2)$ represents a realization of $C(3, 2)$, where $C(3, 2)$ denotes choosing 2 integers from the set (i_1, i_2, i_3) of 3 elements, and let $j_1 < j_2$.

...

$$p_{i_1, i_2, i_3, \dots, i_s}^{(s)} = \text{p(the pool is positive for trait } i_1 \text{ and trait } i_2 \text{ and trait } i_3, \dots \text{ and trait } i_s)$$

$$= \left[\sum \theta_j^{(1)} + \sum \theta_{j_1, j_2}^{(2)} + \sum \theta_{j_1, j_2, j_3}^{(3)} + \dots + \sum \theta_{j_1, j_2, \dots, j_f}^{(f)} + \dots + \theta_{i_1, i_2, \dots, i_s}^{(s)} + \theta^{(0)} \right]^m - p^{(0)} - \sum p_j^{(1)} - \sum p_{j_1, j_2}^{(2)} - \sum p_{j_1, j_2, j_3}^{(3)} - \dots - \sum p_{j_1, j_2, \dots, j_f}^{(f)} - \dots - \sum p_{j_1, j_2, \dots, j_{s-1}}^{(s-1)}$$

where $(j_1, j_2, j_3, \dots, j_f)$ represents a realization of $C(s, f)$, where $C(s, f)$ denotes choosing f integers from the set (i_1, i_2, \dots, i_s) of s elements, and let $j_1 < j_2 < \dots < j_f$.

...

$$p^{(k)} = 1 - p^{(0)} - \sum p_i^{(1)} - \sum p_{i_1, i_2}^{(2)} - \sum p_{i_1, i_2, i_3}^{(3)} - \dots - \sum p_{i_1, i_2, i_3, \dots, i_{k-1}}^{(k-1)}$$

Log-likelihood for a given pattern when n pools are tested for the presence of K traits :

Let $(n^{(0)}, n^{(1)}, n^{(2)}, n^{(3)}, \dots, n^{(s)}, \dots, n^{(k)})$ be the pattern observed for n pools of seeds:

$n^{(0)}$ negative pools,

$n^{(1)}$ positive pools for single traits, a vector of length k

$n^{(2)}$ positive pools for two traits, a vector of length $C_{k,2}$

$n^{(3)}$ positive pools for three traits, a vector of length $C_{k,3}$

...

$n^{(s)}$ positive pools for s traits, a vector of length $C_{k,s}$

...

$n^{(k)}$ positive pools for all K traits, a scalar

$$n = n^{(0)} + n^{(1)} + n^{(2)} + n^{(3)} + \dots + n^{(s)} + \dots + n^{(k)}$$

$$n = n^{(0)} + \sum n^{(1)} + \sum n^{(2)} + \sum n^{(3)} + \dots + \sum n^{(s)} + \dots + n^{(k)}$$

Log-likelihood for a given pattern when n pools are tested for the presence of traits 1, 2, ..., k :

$$l(\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \dots, \theta^{(k)})$$

$$= n^{(0)} \times \ln(p^{(0)}) + (n^{(1)})^T \times \ln(p^{(1)}) + (n^{(2)})^T \times \ln(p^{(2)}) + \dots + (n^{(s)})^T \times \ln(p^{(s)}) + \dots + n^{(k)} \times \ln(p^{(k)})$$

where $()^T$ denotes a vector transpose.



Further theoretical work

Introduction in the formulas of
assay system uncertainty
i.e. **false-positive** and **false-negative** rates





Examples: double stack assessment



- **12** pools of **150** seeds:

Observed pattern:

9 negative pools for A and B

0 positive pools for **A** only

0 positive pools for **B** only

3 positive pools for both **A** and **B**

$$\begin{aligned} \hat{\theta}_A &= 0\% \\ \hat{\theta}_B &= 0\% \\ \hat{\theta}_{AB} &= 0.19\% \end{aligned}$$



- **12** pools of **150** seeds:

Observed pattern:

2 negative pools for A and B

5 positive pools for **A** only

3 positive pools for **B** only

2 positive pools for both **A** and **B**

$$\begin{aligned} \hat{\theta}_A &= 0.58\% \\ \hat{\theta}_B &= 0.36\% \\ \hat{\theta}_{AB} &= 0\% \end{aligned}$$





Examples: double stack assessment

- **10** pools of **150** seeds:

Observed pattern:

7 negative pools for A and B

0 positive pools for **A** only

1 positive pools for **B** only

2 positive pools for both **A** and **B**

$$\hat{\theta}_A = 0\%$$

$$\rightarrow \hat{\theta}_B = 0.09\%$$

$$\hat{\theta}_{AB} = 0.15\%$$

- **4** pools of **1000** seeds:

Observed pattern:

2 negative pools for A and B

0 positive pools for **A** only

1 positive pools for **B** only

1 positive pools for both **A** and **B**

$$\hat{\theta}_A = 0\%$$

$$\rightarrow \hat{\theta}_B = 0.041\%$$

$$\hat{\theta}_{AB} = 0.029\%$$





Examples: triple stack assessment



- **12** pools of **150** seeds:
Observed pattern:
 - 3** negative pools for **A** and **B**
 - 1** positive pools for **A** only
 - 1** positive pools for **B** only
 - 1** positive pools for **C** only
 - 1** positive pools for both **A** and **B**
 - 0** positive pools for both **A** and **C**
 - 2** positive pools for both **B** and **C**
 - 3** positive pools for both **A**, **B** and **C**

$$\begin{aligned} \hat{\theta}_A &= 0.15\% \\ \hat{\theta}_B &= 0.18\% \\ \hat{\theta}_C &= 0.15\% \\ \hat{\theta}_{AB} &= 0.09\% \\ \hat{\theta}_{AC} &= 0\% \\ \hat{\theta}_{BC} &= 0.19\% \\ \hat{\theta}_{ABC} &= 0.12\% \end{aligned}$$



Examples: triple stack assessment



- **15** pools of **250** seeds:
Observed pattern:
 - 8** negative pools for **A** and **B**
 - 0** positive pools for **A** only
 - 0** positive pools for **B** only
 - 1** positive pools for **C** only
 - 0** positive pools for both **A** and **B**
 - 0** positive pools for both **A** and **C**
 - 1** positive pools for both **B** and **C**
 - 5** positive pools for both **A**, **B** and **C**



$$\hat{\theta}_A = 0\%$$

$$\hat{\theta}_B = 0\%$$

$$\hat{\theta}_C = 0.05\%$$

$$\hat{\theta}_{AB} = 0\%$$

$$\hat{\theta}_{AC} = 0\%$$

$$\hat{\theta}_{BC} = 0.04\%$$

$$\hat{\theta}_{ABC} = 0.16\%$$



Implementation

- **R package**

(**R** is a **free software** environment for statistical computing and graphics)

- Less than **1 second** to get solutions for triple stack assessment

- Will be available on **ISTA** website

