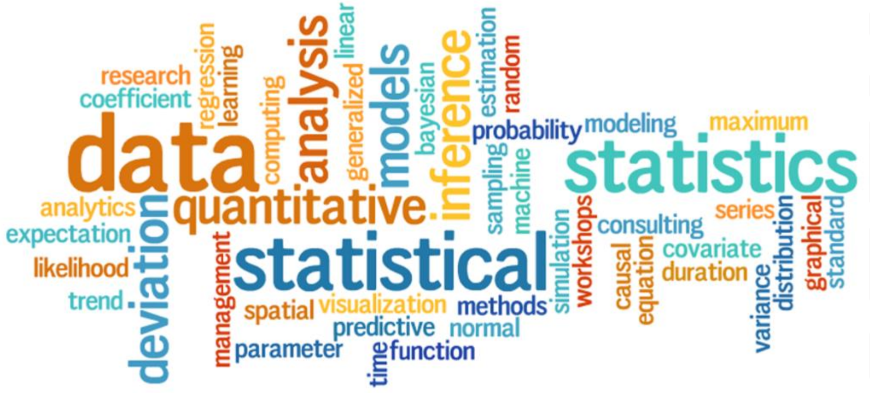
ISTA Statistics Committee open meeting: overview of some recent projects



Presenter: Kirk Remund & Jean-Louis Laffont
Location: Verona, Italy
Date: May 31, 2023

ISTA Statistics Committee

Chair:	Kirk Remund	USA
Vice:	Jean-Louis Laffont	France
Members:	Gabriel Carré	France
	Mustapha El Yakhlifi	France
	Zhou Fang	USA
	Bonnie Hong	USA
	Bo-Jein Kuo	Separate Custom Territory of Taiwan, Penghu, Kinmen and Matsu
	Ray Shillito	USA
	Thomas Michelon	Brasil
	Oluseyi Odubote	USA

Outline

1. New statistical tool for determining working sample weight to amend Table 2C of ISTA Rules

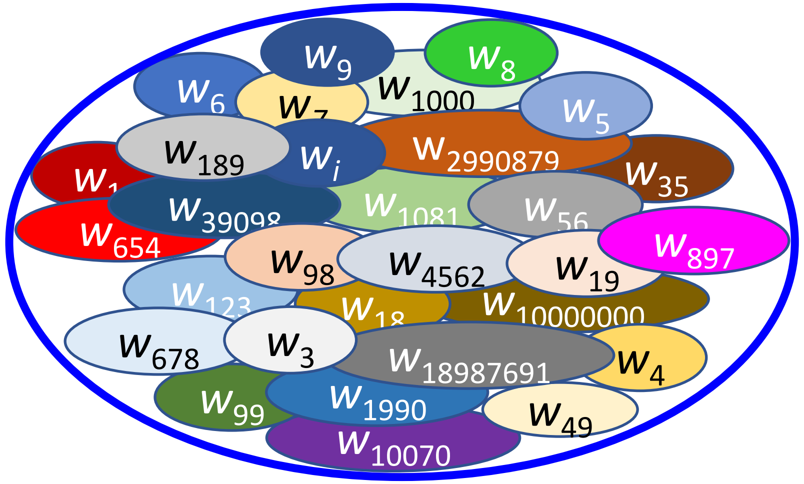
2. Number of sub-lots for which an OIC established for the lot is still valid

3. Group testing: number of groups to ensure that estimation is possible

4. Opportunities…

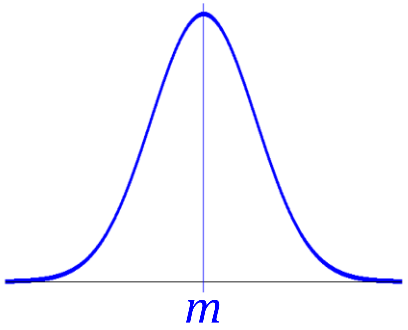
1. New statistical tool for determining working sample weight to amend Table 2C of ISTA Rules

Population (all possible varieties, lots, labs and 100 seeds samples) of 100 seed weights



Y ~ Normal distribution with mean m and variance σ^2

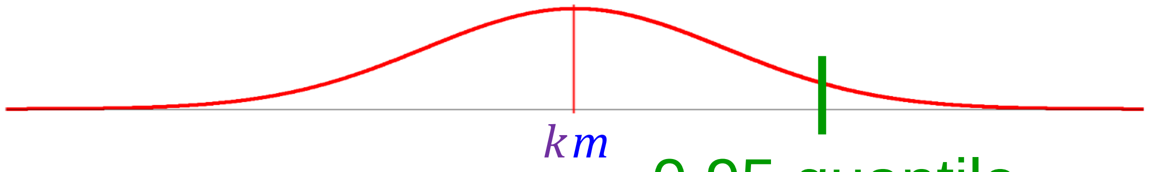
Assumption



X : "2500 (k = 25) or 25000 (k = 250) seeds weight"

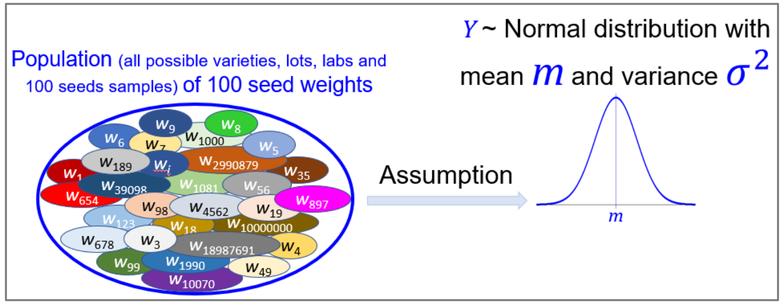
If $Y \sim N(m\ ,\sigma^2)$, then

$X = kY$ is $\sim N(km\ , k^2\sigma^2)$



0.95 quantile

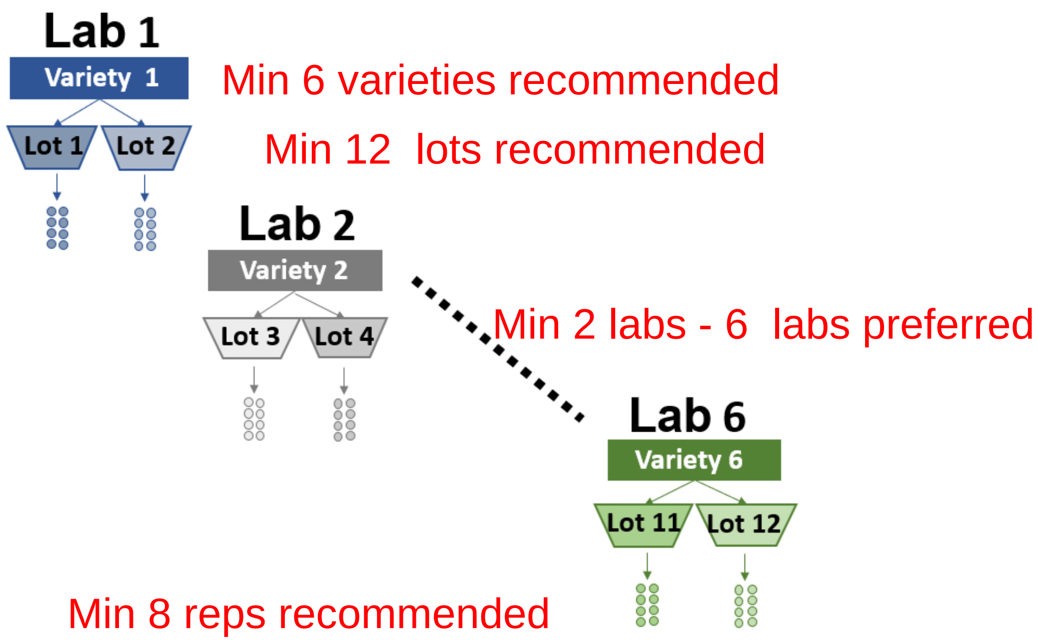
95% confident to have at least 2500 or 25000 seeds in a random sample with the 0.95 quantile weight



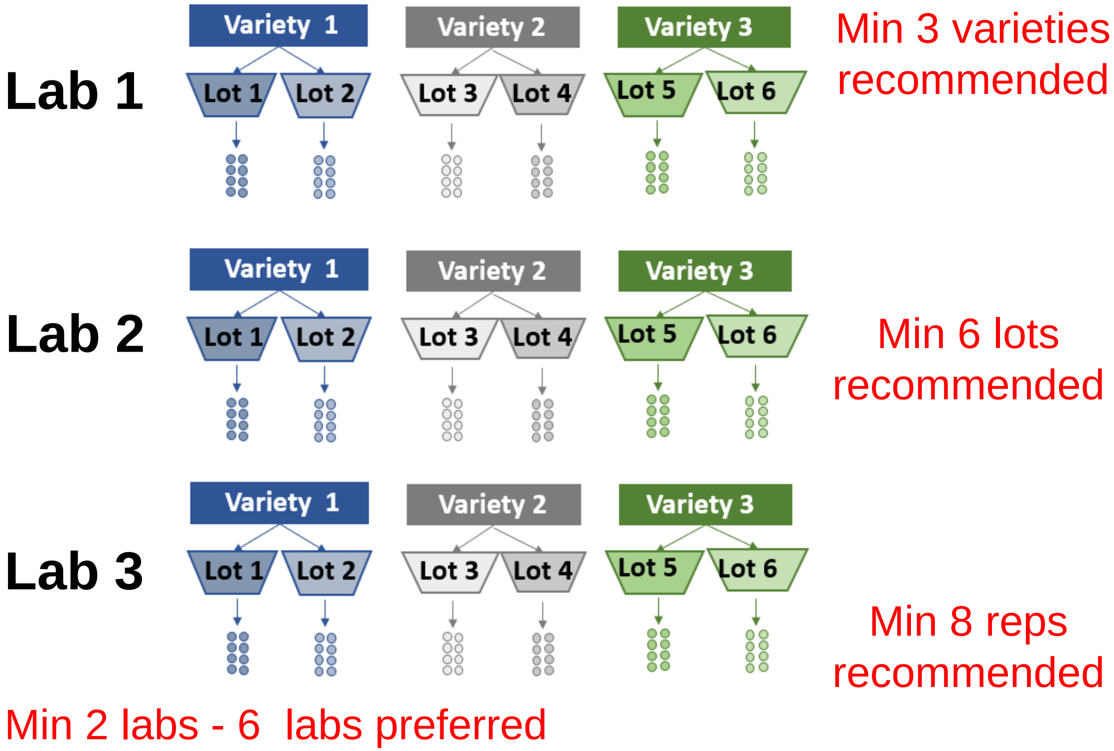
Estimating m and σ^2

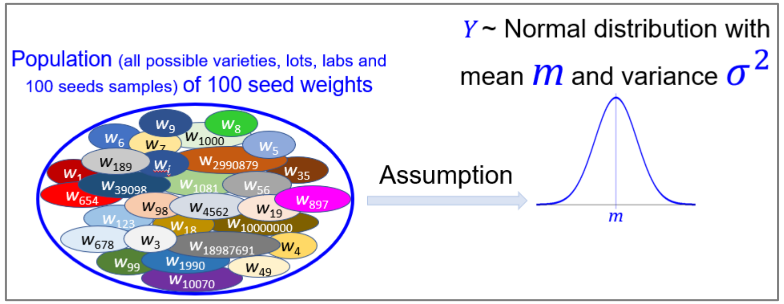
2 experiment designs to capture all the possible sources of variation at its best

Experiment design 1: 2-way nested design



Experiment design 2: 2-way crossed design





Estimating m and σ^2

Fitting linear random effects model

Experiment design 1
(2-way nested design)

$$100\_seeds\_weight = general\_mean$$
$$+ Lab\_effect \rightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lab})$$
$$+ Lot(within\ Lab)\_effect \rightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lot})$$
$$+ Residual \rightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Res})$$

Experiment design 2
(2-way crossed design)

$$100\_seeds\_weight = general\_mean$$
$$+ Lab\_effect \rightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lab})$$
$$+ Lot\_effect \rightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lot})$$
$$+ Lab \times Lot\_effect \rightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lab\times Lot})$$
$$+ Residual \rightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Res})$$

$$\hat{m} = \widehat{general\_mean}$$

$$\widehat{\sigma^2} = \widehat{\sigma^2_{Lab}} + \widehat{\sigma^2_{Lot}} + \widehat{\sigma^2_{Res}}$$

$$\widehat{\sigma^2} = \widehat{\sigma^2_{Lab}} + \widehat{\sigma^2_{Lot}} + \widehat{\sigma^2_{Lab\times Lot}} + \widehat{\sigma^2_{Res}}$$

(estimates are denoted with a "hat")

Prior to estimation, reps outliers are detected using Grubbs's method

i	y_i	T_i
1	0.4460	0.7600748
2	0.4190	0.3052932
3	0.4000	0.0147383
4	0.4270	0.4400433
5	0.2600	2.3728652
6	0.4100	0.1536993
7	0.4420	0.6926998
8	0.4030	0.035793

1. Calculate the mean $\bar{y}$ and the standard-deviation s :

   $\bar{y} = 0.4009$ $\quad\quad$ $s = 0.0594$

2. For each value y_i in the dataset, calculate: $T_i = \dfrac{|y_i - \bar{y}|}{s}$

3. If T_i is greater than a critical value corresponding to a given significance probability (usually 5%), then identify y_i as an outlier

i	y_i	T_i
1	0.4460	0.7600748
2	0.4190	0.3052932
3	0.4000	0.0147383
4	0.4270	0.4400433
5	0.2600	2.3728652
6	0.4100	0.1536993
7	0.4420	0.6926998
8	0.4030	0.035793

Critical values for 5% level of significance

Sample size	Critical value	Sample size	Critical value	Sample size	Critical value	Sample size	Critical value
3	1.15	15	2.55	27	2.86	39	3.03
4	1.48	16	2.59	28	2.88	40	3.04
5	1.71	17	2.62	29	2.89	50	3.13
6	1.89	18	2.65	30	2.91	60	3.20
7	2.02	19	2.68	31	2.92	70	3.26
8	2.13	20	2.71	32	2.94	80	3.31
9	2.21	21	2.73	33	2.95	90	3.35
10	2.29	22	2.76	34	2.97	100	3.38
11	2.34	23	2.78	35	2.98	110	3.42
12	2.41	24	2.80	36	2.99	120	3.44
13	2.46	25	2.82	37	3.00	130	3.47
14	2.51	26	2.84	38	3.01	140	3.49

→ y_5 is identified as an outlier

Grubbs's method critical values can be calculated from the student distribution as follows:

$$\sqrt{\frac{\left[(n-1)t_{1-\frac{\alpha}{2n},n-2}\right]^2}{n\left[(n-2)+\left(t_{1-\frac{\alpha}{2n},n-2}\right)^2\right]}}$$

where:
- n : sample size
- α : level of significance
- $t_{1-\frac{\alpha}{2n},n-2}$: $1-\frac{\alpha}{2n}$ critical point of a t-distribution with $n-2$ degrees of freedom

Can be easily implemented into Excel:

B6   fx   `=SQRT(((($B$4-1)*TINV($B$3/$B$4,$B$4-2))^2)/($B$4*(($B$4-2)+TINV($B$3/$B$4,$B$4-2)^2)))`

	A	B	C	D	E	F	G
1	**Grubbs' method: critical values**						
2							
3	**Level of significance:**	**5%**					
4	**Sample size:**	**15**					
5							
6	**Grubbs' critical value:**	**2.55**					
7							

2-way nested design

$$
\begin{aligned}
100\_seeds\_weight = {} & general\_mean \\
& + Lab\_effect \longrightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lab}) \\
& + Lot(within\ Lab)\_effect \longrightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lot}) \\
& + Residual \longrightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Res})
\end{aligned}
$$

2-way crossed design

$$
\begin{aligned}
100\_seeds\_weight = {} & general\_mean \\
& + Lab\_effect \longrightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lab}) \\
& + Lot\_effect \longrightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lot}) \\
& + Lab \times Lot\_effect \longrightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Lab \times Lot}) \\
& + Residual \longrightarrow \sim \text{i.i.d. } N(0,\sigma^2_{Res})
\end{aligned}
$$

- There are several methods to get estimates of variance components:

  . ANOVA based methods

  . **M**aximum **L**ikelihood (**ML**) methods

  . **RE**stricted **M**aximum **L**ikelihood (**REML**) methods

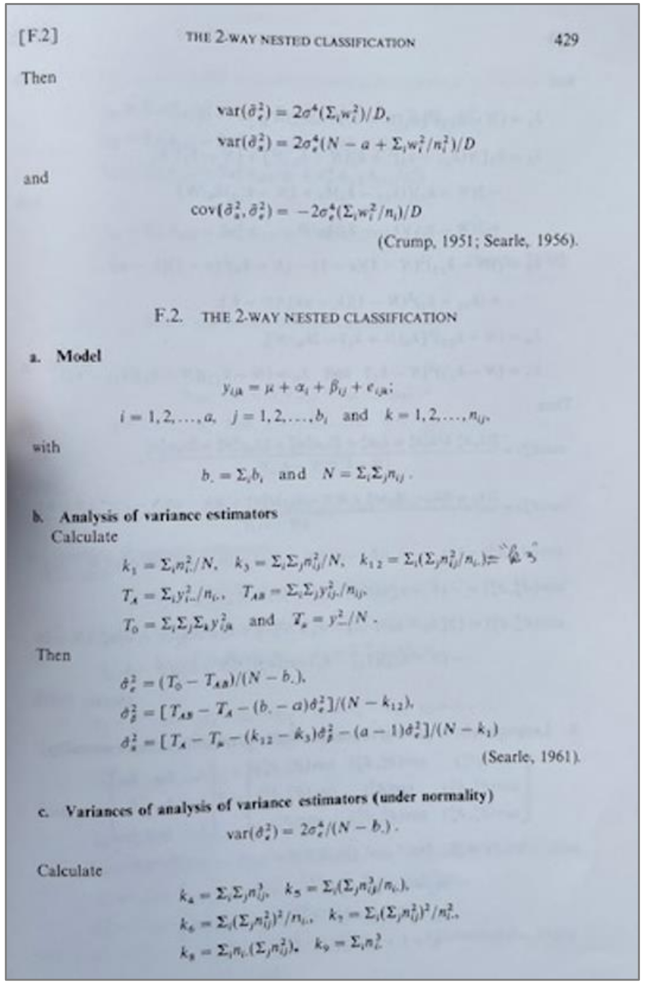
  …
- Today, the preferred method is REML … but it requires heavy computations

⟹ Selected **Henderson Method I** (ANOVA based method) for its ease of implementation in Excel.
This method works for unbalanced data; for balanced data, it provides identical estimates as REML method

Henderson Method I

Searle, S.R., Casella, G. and C.E. McCulloch (1992).
In Variance components (pp. 429, 434-435). *Wiley-Interscience*, New York.

This is what is implemented in the calculator



[F.2] THE 2-WAY NESTED CLASSIFICATION 429

Then

$$\mathrm{var}(\hat\sigma_e^2) = 2\sigma^4(\Sigma_i w_i^2)/D,$$
$$\mathrm{var}(\hat\sigma_\alpha^2) = 2\sigma_e^4(N - a + \Sigma_i w_i^2/n_i^2)/D$$

and

$$\mathrm{cov}(\hat\sigma_\alpha^2, \hat\sigma_e^2) = -2\sigma_e^4(\Sigma_i w_i^2/n_i)/D$$

(Crump, 1951; Searle, 1956).

F.2. THE 2-WAY NESTED CLASSIFICATION

a. Model

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk};$$
$$i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b_i \quad \text{and} \quad k = 1, 2, \ldots, n_{ij},$$

with

$$b_. = \Sigma_i b_i \quad \text{and} \quad N = \Sigma_i \Sigma_j n_{ij}.$$

b. Analysis of variance estimators
Calculate

$$k_1 = \Sigma_i n_{i.}^2/N, \quad k_3 = \Sigma_i\Sigma_j n_{ij}^2/N, \quad k_{12} = \Sigma_i(\Sigma_j n_{ij}^2/n_{i.}),$$
$$T_A = \Sigma_i y_{i..}^2/n_{i.}, \quad T_{AB} = \Sigma_i\Sigma_j y_{ij.}^2/n_{ij},$$
$$T_0 = \Sigma_i\Sigma_j\Sigma_k y_{ijk}^2 \quad \text{and} \quad T_\mu = y_{...}^2/N.$$

Then

$$\hat\sigma_e^2 = (T_0 - T_{AB})/(N - b_.),$$
$$\hat\sigma_\beta^2 = [T_{AB} - T_A - (b_. - a)\hat\sigma_e^2]/(N - k_{12}),$$
$$\hat\sigma_\alpha^2 = [T_A - T_\mu - (k_{12} - k_3)\hat\sigma_\beta^2 - (a-1)\hat\sigma_e^2]/(N - k_1)$$

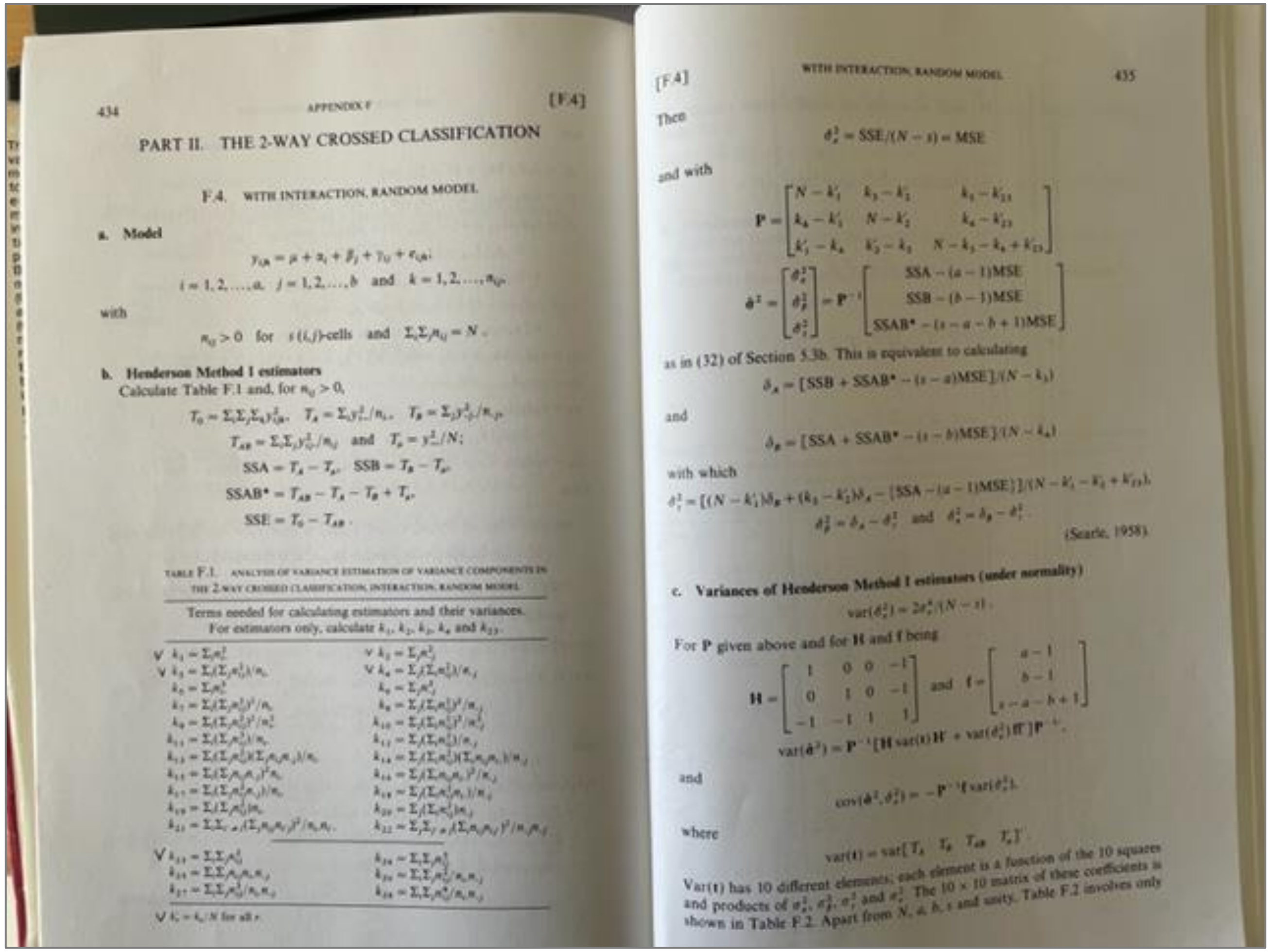
(Searle, 1961).

c. Variances of analysis of variance estimators (under normality)

$$\mathrm{var}(\hat\sigma_e^2) = 2\sigma_e^4/(N - b_.).$$

Calculate

$$k_4 = \Sigma_i\Sigma_j n_{ij}^3, \quad k_5 = \Sigma_i(\Sigma_j n_{ij}^3/n_{i.}),$$
$$k_6 = \Sigma_i(\Sigma_j n_{ij}^2)^2/n_{i.}, \quad k_7 = \Sigma_i(\Sigma_j n_{ij}^2)^2/n_{i.}^2,$$
$$k_8 = \Sigma_i n_{i.}(\Sigma_j n_{ij}^2), \quad k_9 = \Sigma_i n_{i.}^3$$



434 APPENDIX F [F.4]

PART II. THE 2-WAY CROSSED CLASSIFICATION

F.4. WITH INTERACTION, RANDOM MODEL

a. Model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk};$$
$$i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b \quad \text{and} \quad k = 1, 2, \ldots, n_{ij},$$

with

$$n_{ij} > 0 \text{ for } s\ (i,j)\text{-cells and } \Sigma_i\Sigma_j n_{ij} = N.$$

b. Henderson Method I estimators
Calculate Table F.1 and, for $n_{ij} > 0$,

$$T_0 = \Sigma_i\Sigma_j\Sigma_k y_{ijk}^2, \quad T_A = \Sigma_i y_{i..}^2/n_{i.}, \quad T_B = \Sigma_j y_{.j.}^2/n_{.j},$$
$$T_{AB} = \Sigma_i\Sigma_j y_{ij.}^2/n_{ij} \quad \text{and} \quad T_\mu = y_{...}^2/N;$$
$$\mathrm{SSA} = T_A - T_\mu, \quad \mathrm{SSB} = T_B - T_\mu,$$
$$\mathrm{SSAB}^* = T_{AB} - T_A - T_B + T_\mu,$$
$$\mathrm{SSE} = T_0 - T_{AB}.$$

TABLE F.1. ANALYSIS OF VARIANCE ESTIMATION OF VARIANCE COMPONENTS IN THE 2-WAY CROSSED CLASSIFICATION, INTERACTION, RANDOM MODEL

Terms needed for calculating estimators and their variances.
For estimators only, calculate k_1, k_2, k_3, k_4 and k_{23}.

[F.4] WITH INTERACTION, RANDOM MODEL 435

Then

$$\hat\sigma_e^2 = \mathrm{SSE}/(N - s) = \mathrm{MSE}$$

and with

$$\mathbf{P} = \begin{bmatrix} N - k_1' & k_3 - k_2' & k_3 - k_{23}' \\ k_4 - k_1' & N - k_2' & k_4 - k_{23}' \\ k_1' - k_4 & k_2' - k_3 & N - k_3 - k_4 + k_{23}' \end{bmatrix}$$

$$\hat{\boldsymbol\sigma}^2 = \begin{bmatrix} \hat\sigma_\alpha^2 \\ \hat\sigma_\beta^2 \\ \hat\sigma_\gamma^2 \end{bmatrix} = \mathbf{P}^{-1}\begin{bmatrix} \mathrm{SSA} - (a-1)\mathrm{MSE} \\ \mathrm{SSB} - (b-1)\mathrm{MSE} \\ \mathrm{SSAB}^* - (s - a - b + 1)\mathrm{MSE} \end{bmatrix}$$

as in (32) of Section 5.3b. This is equivalent to calculating

$$\delta_A = [\mathrm{SSB} + \mathrm{SSAB}^* - (s-a)\mathrm{MSE}]/(N - k_3)$$

and

$$\delta_B = [\mathrm{SSA} + \mathrm{SSAB}^* - (s-b)\mathrm{MSE}]/(N - k_4)$$

with which

$$\hat\sigma_\gamma^2 = [(N - k_1')\delta_B + (k_3 - k_2')\delta_A - \{\mathrm{SSA} - (a-1)\mathrm{MSE}\}]/(N - k_1' - k_2' + k_{23}'),$$
$$\hat\sigma_\beta^2 = \delta_A - \hat\sigma_\gamma^2 \quad \text{and} \quad \hat\sigma_\alpha^2 = \delta_B - \hat\sigma_\gamma^2$$

(Searle, 1958).

c. Variances of Henderson Method I estimators (under normality)

$$\mathrm{var}(\hat\sigma_e^2) = 2\sigma_e^4/(N - s).$$

For $\mathbf{P}$ given above and for $\mathbf{H}$ and $\mathbf{f}$ being

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} a-1 \\ b-1 \\ s-a-b+1 \end{bmatrix}$$

$$\mathrm{var}(\hat{\boldsymbol\sigma}^2) = \mathbf{P}^{-1}[\mathbf{H}\,\mathrm{var}(\mathbf{t})\mathbf{H}' + \mathrm{var}(\hat\sigma_e^2)\mathbf{f}\mathbf{f}']\mathbf{P}^{-1\prime}$$

and

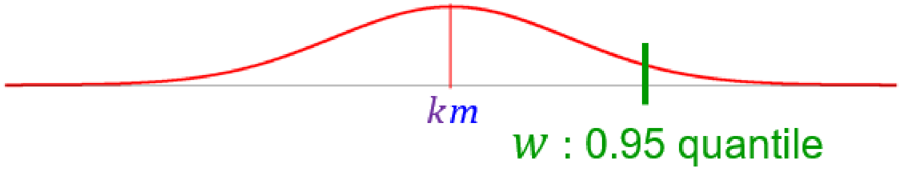
$$\mathrm{cov}(\hat{\boldsymbol\sigma}^2, \hat\sigma_e^2) = -\mathbf{P}^{-1}\mathbf{f}\,\mathrm{var}(\hat\sigma_e^2),$$

where

$$\mathrm{var}(\mathbf{t}) = \mathrm{var}[T_A \; T_B \; T_{AB} \; T_\mu]'.$$

Var(t) has 10 different elements; each element is a function of the 10 squares and products of σ_α^2, σ_β^2, σ_γ^2 and σ_e^2. The 10 × 10 matrix of these coefficients is shown in Table F.2. Apart from N, a, b, s and unity, Table F.2 involves only

X : "2500 (k = 25) or 25000 (k = 250) seeds weight"



Reporting

- If $w < 1\,\mathrm{g}$, w is rounded up to the nearest multiple of 0.01
- If $1\,\mathrm{g} \leq w < 5\,\mathrm{g}$, w is rounded up to the nearest multiple of 0.1
- If $w \geq 5\,\mathrm{g}$, w is rounded up to the nearest integer

Examples:

w	Value reported
0.34567	0.35
0.96781	0.97
0.99001	1.00
1.08962	1.1
4.45687	4.5
5.00768	6
9.76981	10

1. Overview of the calculator

Instructions | Calculator

Calculator for adding working weights to Table 2C of the ISTA Rules

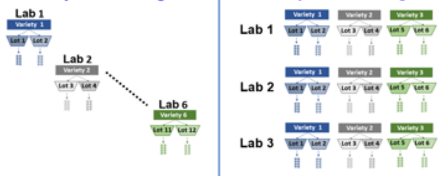
THE CACULATOR IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND. IN NO EVENT SHALL THE AUTHORS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY ARISING IN CONNECTION WITH THE CALCULATOR.

Experiment designs

Two types of experiment designs are considered in the calculator:

Experiment design 1: 2-way nested design

Experiment design 2: 2-way crossed design



A minimum of 12 lots are considered across a minimum of six varieties represented in the experiment as a general rule.
These 12 lots will be evaluated by a minimum of two labs however six labs are preferred.
A minimum of eight 100 seed reps are weighed per lot.

A minimum of six lots are considered across a minimum of three varieties represented in the experiment as a general rule.
These six lots will be evaluated each by a minimum of two labs however six labs are preferred.
A minimum of eight 100 seed reps are weighed per lot.

Calculations

The rep weights are entered into the unprotected yellow cells of the calculator. If experiment design 1 is used, data for the different lots from each lab are entered in different columns. In order to avoid conditional formatting conflicts, always copy/paste data in the calculator using `Paste Special → Values`.

. For each lot in a given laboratory, outliers are highlighted in red using Grubbs's method at the 5% significance level (Grubbs, 1969). These outliers are then excluded manually from the computations.

. The linear random effects models used for the analysis of the two experiment designs are:

. **Experiment design 1**:

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}$$

in which:
. y_{ijk} is the observed 100-seeds weight of lot j ($j = 1, 2, ..., b_i$) in lab i ($i = 1, 2, ..., a$) and replication k ($k = 1, 2, ..., n_{ij}$);
. μ is the intercept;
. α_i is the random effect of lab i ($\alpha_i \sim \text{i.i.d.}\, N(0, \sigma^2_{Lab})$);
. β_{ij} is the random effect of lot j within lab i ($\beta_{ij} \sim \text{i.i.d.}\, N(0, \sigma^2_{Lot})$);
. e_{ijk} is the residual ($e_{ijk} \sim \text{i.i.d.}\, N(0, \sigma^2_{Res})$).

. **Experiment design 2**:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

in which:
. y_{ijk} is the observed 100-seeds weight of lot j ($j = 1, 2, ..., b$) in lab i ($i = 1, 2, ..., a$) and replication k ($k = 1, 2, ..., n_{ij}$);
. μ is the intercept;
. α_i is the random effect of lab i ($\alpha_i \sim \text{i.i.d.}\, N(0, \sigma^2_{Lab})$);
. β_j is the random effect of lot j ($\beta_j \sim \text{i.i.d.}\, N(0, \sigma^2_{Lot})$);
. $(\alpha\beta)_{ij}$ is the random interaction effect between lab i and lot j ($(\alpha\beta)_{ij} \sim \text{i.i.d.}\, N(0, \sigma^2_{Lab\times Lot})$);
. e_{ijk} is the residual ($e_{ijk} \sim \text{i.i.d.}\, N(0, \sigma^2_{Res})$).

The calculator automatically selects which model to fit according to the dataset structure.

. Variance components for the two models are estimated from the data by the Henderson Method I (Searle *et al.*, 1992, Appendix F). When an estimate is negative, this estimate is reported as zero. Let $\hat{\sigma}^2_{Lab}$, $\hat{\sigma}^2_{Lot}$, $\hat{\sigma}^2_{Lab\times Lot}$ and $\hat{\sigma}^2_{Res}$ be these

Supporting Data of New Species Proposal to ISTA Rules Table 2C

Submitter Name:		Lab Full Name:	
Scientific Name of the Crop kind:	Genus Species	ISTA Member Code:	
		Contact Email:	

Change any value in a yellow cell

Number of observations	0	
Number of labs	0	
Number of lots	0	
General mean		
Lab variance		
Lot variance		
Lab x Lot variance		
Residual variance		Decision
2500 seed weight*		
25000 seed weight*		

* 95% Confidence

Rep weights in red are identified as outliers by Grubbs's method at the 5% significance level and needs to be suppressed (removed) manually

Lab	\ Seed lot	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Lab 1	Rep1																	
	Rep2																	
	Rep3																	
	Rep4																	
	Rep5																	
	Rep6																	
	Rep7																	
	Rep8																	
	Rep9																	
	Rep10																	
	Rep11																	
	Rep12																	
	Mean																	
	St. Dev.																	
	Number of reps	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Grubbs critical values																	
Lab 2	Rep1																	
	Rep2																	
	Rep3																	
	Rep4																	
	Rep5																	
	Rep6																	
	Rep7																	
	Rep8																	
	Rep9																	
	Rep10																	
	Rep11																	

Instructions | Calculator

- The spreadsheet is protected (no password): entering data is only possible in yellow cells
- In order to avoid conditional formatting conflicts, always copy/paste data in the calculator using Paste Special → Values

When needed, some warnings are displayed in red

Supporting Data of New Species Proposal to ISTA Rules Table 2C

Submitter Name:	XXX	Lab Full Name:	YYY
Scientific Name of the Crop kind:	Basella B. alba	ISTA Member Code:	ZZZ
		Contact Email:	AAA

Change any value in a yellow cell

Number of observations	232		
Number of labs	7		
Number of lots	5		6 lots are preferred for an accurate estimation
General mean	3.2584		
Lab variance	0.0020037		
Lot variance	0.1453077		
Lab x Lot variance	0.0963083		
Residual variance	0.0101622	Decision	
2500 seed weight*	103	100	Decision value should be greater than or equal to 103
25000 seed weight*	1022	1050	

* 95% Confidence

Rep weights in red are identified as outliers by Grubbs's method at the 5% significance level and needs to be suppressed (removed) manually

Lab	\ Seed lot	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Lab 2	Rep1	2.3672	3.4036	2.3585	3.1927	3.7473											
	Rep2	2.2734	3.4207	2.3530	3.0972	3.7309											
	Rep3	2.3198	3.5878	2.4268	3.2861	3.7818											
	Rep4	2.3866	3.4322	2.3827	3.2858	3.7380											
	Rep5	2.3600	3.3296	2.2917	3.2861	3.7908											
	Rep6	2.3720	3.3873	2.2663	3.1889	3.8542											
	Rep7		3.4601	2.2407	3.1103												
	Rep8			2.7779	3.2201												
	Rep9																
	Rep10																
	Rep11																
	Rep12																
	Mean	2.3465	3.4316	2.3872	3.2084	3.7738											
	St. Dev.	0.04225	0.08008	0.16963	0.07636	0.04613											
	Number of reps	6	7	8	8	6	0	0	0	0	0	0	0	0	0	0	0
	Grubbs critical values	1.89	2.02	2.13	2.13	1.89											

Outliers are automatically identified in red

Although not recommended, the calculator can provide estimates of the variance components when there is only 1 lab

Random effects model: $100\_seeds\_weight = general\_mean + Lot\_effect + Residual$

Lot_effect ~ i.i.d. $N(0,\sigma^2_{Lot})$

$Residual$ ~ i.i.d. $N(0,\sigma^2_{Res})$

Balanced dataset example:

Number of observations	224	
Number of labs	1	
Number of lots	28	
General mean	0.4261	
~~Lab variance~~		
Lot variance	0.0040998	
~~Lab x Lot variance~~		
Residual variance	0.0003288	Decision
2500 seed weight*	14	
25000 seed weight*	134	

* 95% Confidence

6 labs are preferred for an accurate estimation

```
REML estimates (R package lme4)

Number of obs: 224, groups:  Lot, 28

Random effects:
 Groups   Variance
 Lot      0.0040998
 Residual 0.0003288
```

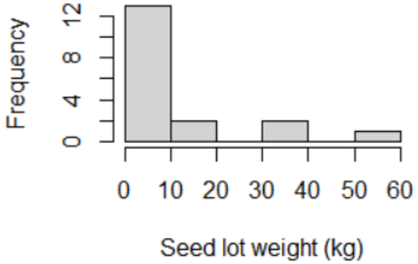
2. Number of sub-lots for which an OIC established for the lot is still valid

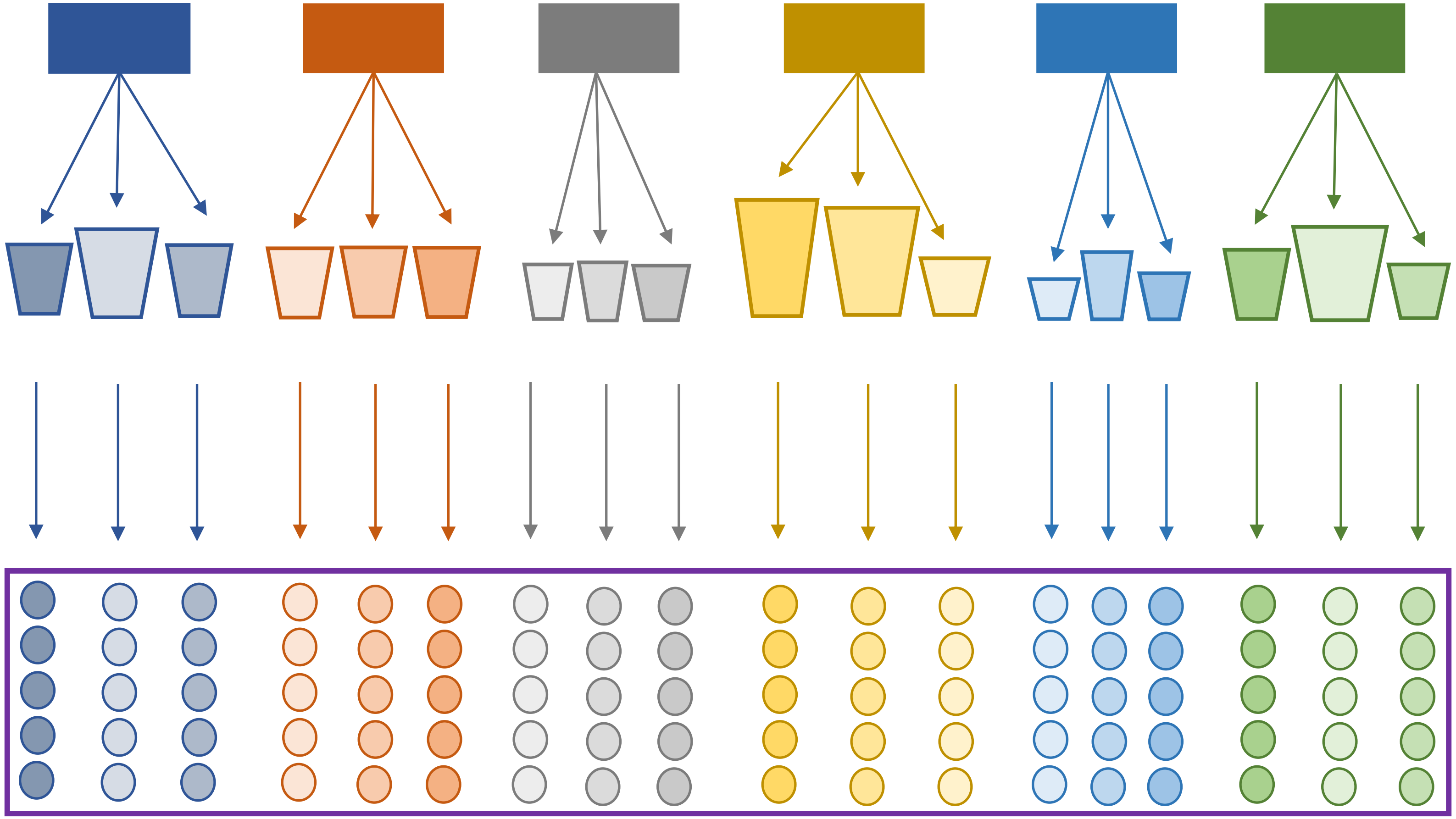
2. 2021 Tomato experiment – Experiment design

6 companies

3 seed lots/company produced in **11** different countries

Different sizes:





90 samples

5 sub-samples taken to cover spatially across each lot

Question:
are the **5** *sub-samples results homogeneous?*

Measurements:

- Purity test
- Germination test: 1st count at day 6, final count at days 8 to 14

Purity % are all equal to 100%

1. Homogeneity of the test replications for normal seedlings, final count:

   → for each of the **90** samples, the **4** reps are within ISTA tolerances

2. Heterogeneity of the **5** samples from each lot

   → Use of the *H* statistic:

$$H = \frac{\overset{400}{\#\_of\_seeds\_in\_the\_sample} \times (\overset{5}{\#\_of\_subsamples} - 1) \times observed\_subsample\_variance}{mean \times (100 - mean)}$$

   → *H* has a chi-squared distribution

   → Statistical test: p-value that all the sample values are equal

   → The lower the p-value, the greater the statistical evidence for heterogeneity

2. 2021 Tomato experiment – Analysis – Sub-samples homogeneity for germination

Company	Lot weight	Normal seedlings %, 1st count			Normal seedlings %, final count			Abnormal seedlings %			Dead seeds %		
		Mean	*H*	p-value	Mean	*H*	p-value	Mean	*H*	p-value	Mean	*H*	p-value
A	3kg	85.8	8.80	0.0663	96.0	4.17	0.3839	3.6	1.38	0.8471	0.4	12.05	0.0170
A	5.9kg	76.6	18.57	0.0010	92.0	2.17	0.7038	3.0	2.75	0.6006	5.0	1.68	0.7936
A	7.7kg	84.8	17.01	0.0019	89.0	10.62	0.0311	8.2	2.55	0.6356	2.8	15.87	0.0032
B	1.5kg	87.2	5.30	0.2575	89.6	1.37	0.8488	3.0	0.00	1.0000	7.4	1.87	0.7600
B	20kg	78.2	16.61	0.0023	98.2	1.81	0.7706	1.8	1.81	0.7706	0.0		
B	6kg	63.8	12.26	0.0155	92.0	5.43	0.2455	4.6	8.39	0.0784	3.4	3.90	0.4201
C	5.9kg	42.0	46.63	0.0000	98.6	3.48	0.4813	1.2	2.70	0.6094	0.2	16.03	0.0030
C	6.4kg	34.8	18.48	0.0010	89.6	4.81	0.3076	7.6	2.96	0.5642	2.8	4.12	0.3906
C	7.8kg	60.6	41.08	0.0000	98.4	8.13	0.0869	1.0	8.08	0.0887	0.6	8.05	0.0898
D	2.1kg	71.4	18.26	0.0011	96.4	3.69	0.4498	1.6	3.05	0.5497	2.0	4.08	0.3951
D	3.3kg	96.0	27.08	0.0000	98.2	6.34	0.1754	1.0	8.08	0.0887	0.8	4.03	0.4017
D	3.7kg	17.8	29.74	0.0000	96.8	6.20	0.1848	1.8	6.34	0.1754	1.4	3.48	0.4813
E	13kg	37.8	29.74	0.0000	97.4	5.05	0.2818	1.6	8.13	0.0869	1.0	8.08	0.0887
E	6.8kg	90.2	8.51	0.0747	96.4	1.38	0.8471	1.4	3.48	0.4813	2.2	5.21	0.2669
E	7.8kg	74.6	30.23	0.0000	98.4	3.05	0.5497	1.2	2.70	0.6094	0.4	12.05	0.0170
F	32kg	84.6	18.79	0.0009	94.2	18.16	0.0012	1.2	2.70	0.6094	3.4	23.38	0.0001
F	36kg	90.2	3.98	0.4084	95.8	6.76	0.1491	2.0	4.08	0.3951	2.2	5.21	0.2669
F	51kg	97.0	5.50	0.2399	97.0	5.50	0.2399	1.6	3.05	0.5497	1.4	3.48	0.4813

Evidence for **heterogeneity**

Evidence for **homogeneity**

- Homogeneity for the germination final count is reinforced by the R test:

		Normal seedlings %, final count				
Company	Lot weight	Mean	*H*	p-value H test	Range	p-value R test
A	3kg	96.0	4.17	0.3839	2.0	0.5995
A	5.9kg	92.0	2.17	0.7038	1.5	0.8355
A	7.7kg	89.0	10.62	0.0311	3.8	0.0522
B	1.5kg	89.6	1.37	0.8488	1.3	0.8867
B	20kg	98.2	1.81	0.7706	1.5	0.8251
B	6kg	92.0	5.43	0.2455	2.9	0.2264
C	5.9kg	98.6	3.48	0.4813	1.7	0.7493
C	6.4kg	89.6	4.81	0.3076	2.6	0.3430
C	7.8kg	98.4	8.13	0.0869	3.2	0.1601
D	2.1kg	96.4	3.69	0.4498	2.1	0.5505
D	3.3kg	98.2	6.34	0.1754	3.0	0.2083
D	3.7kg	96.8	6.20	0.1848	3.4	0.1124
E	13kg	97.4	5.05	0.2818	2.5	0.3868
E	6.8kg	96.4	1.38	0.8471	1.1	0.9422
E	7.8kg	98.4	3.05	0.5497	1.6	0.7922
F	32kg	94.2	18.16	0.0012	6.0	0.0002
F	36kg	95.8	6.76	0.1491	3.0	0.2135
F	51kg	97.0	5.50	0.2399	2.3	0.4602

2. Number of sub-lots determination – Overview

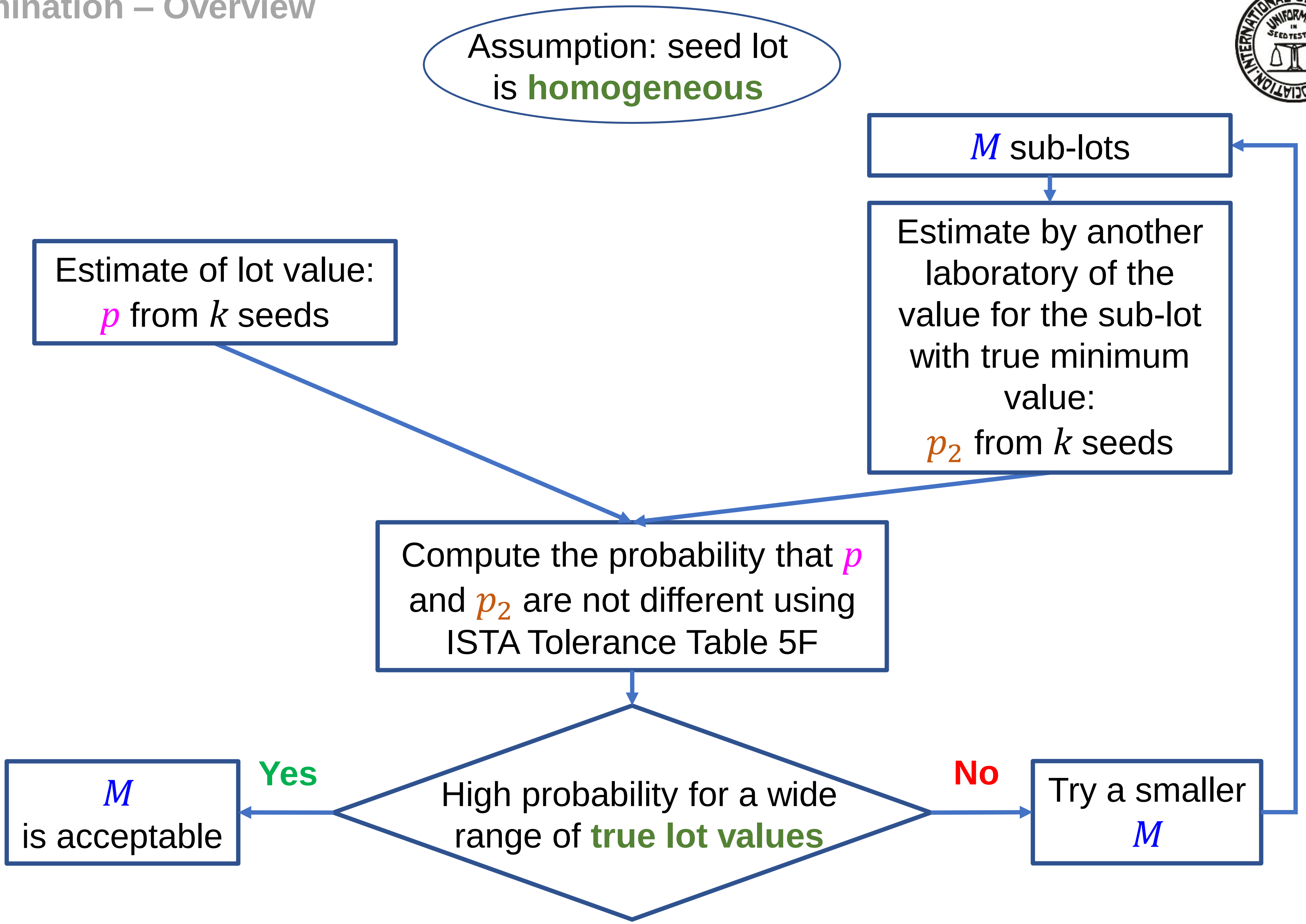


Table 5F. Tolerances between results of two tests made in different laboratories on the same or different samples from the same seed lot (two-way test at 5 % significance level) on 400 seed tests. Updated by ISTA Statistics Technical Committee, based on Miles (1963) Table G5, column C, 400 seed tests.

Average germination percentage of 2 tests		Tolerance
51–100 %	0–50 %	
99	2	2
98	3	3
96–97	4–5	4
94–95	6–7	5
91–93	8–10	6
88–90	11–13	7
84–87	14–17	8
79–83	18–22	9
74–78	23–27	10
68–73	28–33	11
60–67	34–41	12
51–59	42–50	13

2. Number of sub-lots determination – Some details

Assumption: seed lot is **homogeneous**

True lot value: π

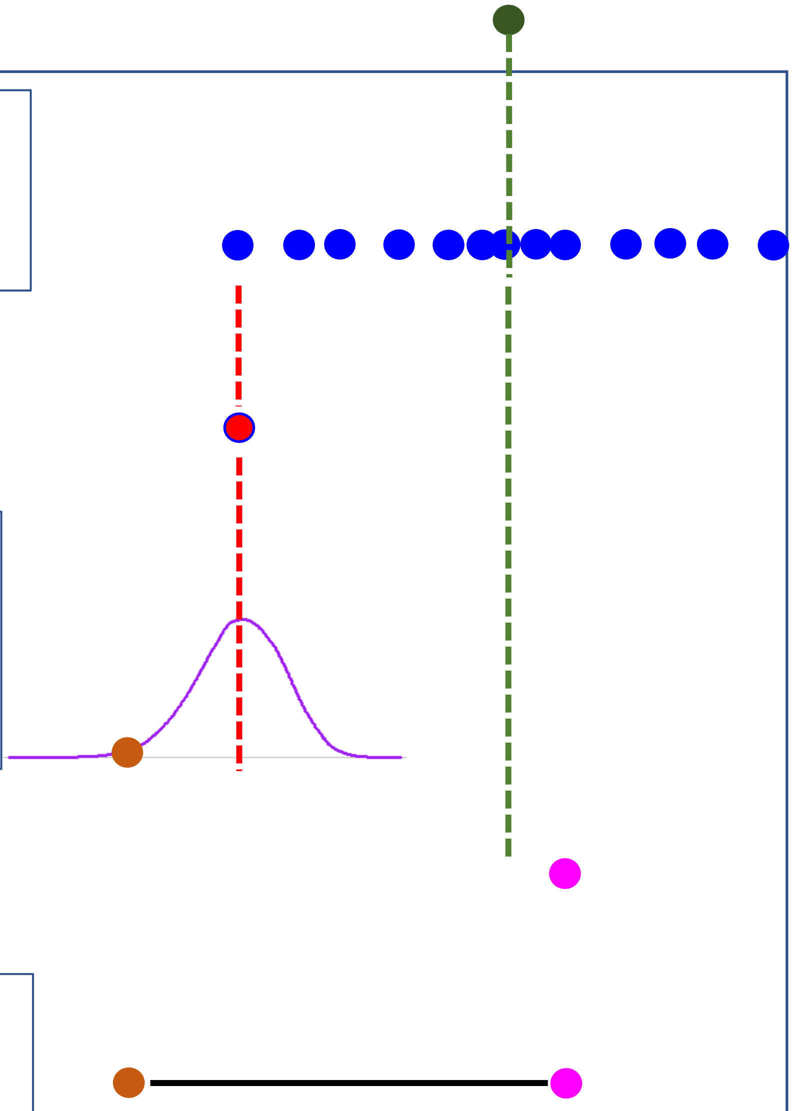
M true sub-lot values π_i from multivariate hypergeometric distribution *

Minimum of the true M sub-lot values: $\pi_m = min\left(\{\pi_i\}_{i=1}^{M}\right)$

Test value from another lab on the sub-lot with minimum value from a **Beta-binomial** distribution: p_2

Estimate of lot value: p from k seeds

Determine if p and p_2 are within tolerance according to ISTA Tolerance Table 5F



Repeated 10,000 times

Compute the probability that p is not different from p_2 as $\frac{T}{10000}$ where T is the number of times p and p_2 are within tolerance

Table 5F. Tolerances between results of two tests made in different laboratories on the same or different samples from the same seed lot (two-way test at 5 % significance level) on 400 seed tests. Updated by ISTA Statistics Technical Committee, based on Miles (1963) Table G5, column C, 400 seed tests.

Average germination percentage of 2 tests		Tolerance
51–100 %	0–50 %	
99	2	2
98	3	3
96–97	4–5	4
94–95	6–7	5
91–93	8–10	6
88–90	11–13	7
84–87	14–17	8
79–83	18–22	9
74–78	23–27	10
68–73	28–33	11
60–67	34–41	12
51–59	42–50	13

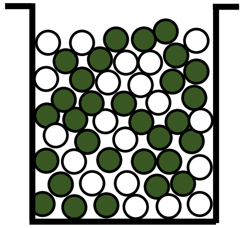
* Laffont, J-L., Hong, B., Kuo, B-J. and K.M. Remund (2019). Exact theoretical distributions around the replicate results of a germination test. *Seed Science Research* **29**, 64-72.

Seed Science Research
cambridge.org/ssr

Exact theoretical distributions around the replicate results of a germination test

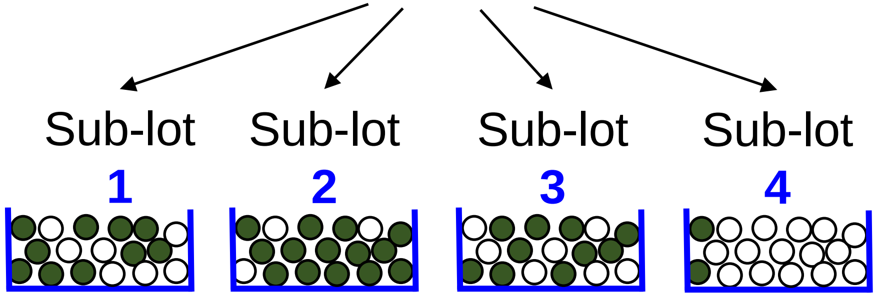
Jean-Louis Laffont[1], Bonnie Hong[2], Bo-Jein Kuo[3] and Kirk M. Remund[4]

Seed lot: N seeds with
$G = N\pi$ seeds to germinate



N = 1,600,000
π = 90%
G = 1,440,000

Random assignment of the seeds into M = 4 sub-lots (n = 400,000 seeds per sub-lot)

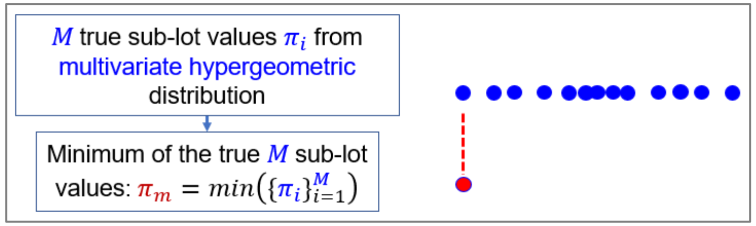


# of seeds to germinate in each sub-lot

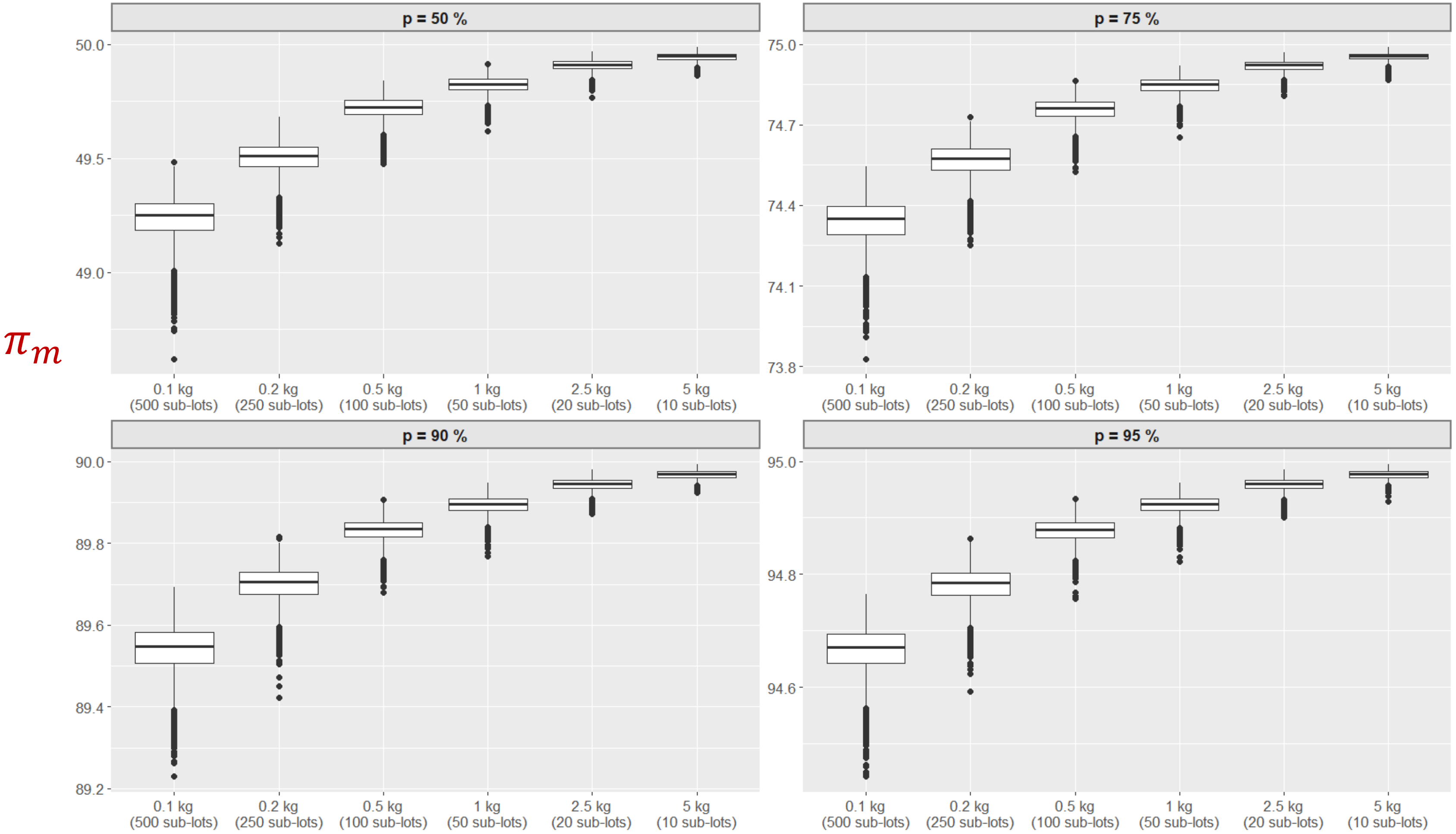
360,023	**360,007**	**359,968**	**360,002**
359,270	**312,987**	**368,976**	**398,767**
360,030	**292,345**	**387,969**	**399,656**
…			

Distribution of the number of seeds to germinate in the sub-lots: **multivariate hypergeometric** distribution

2. Number of sub-lots determination – Some details



Lot weight = 50 kg



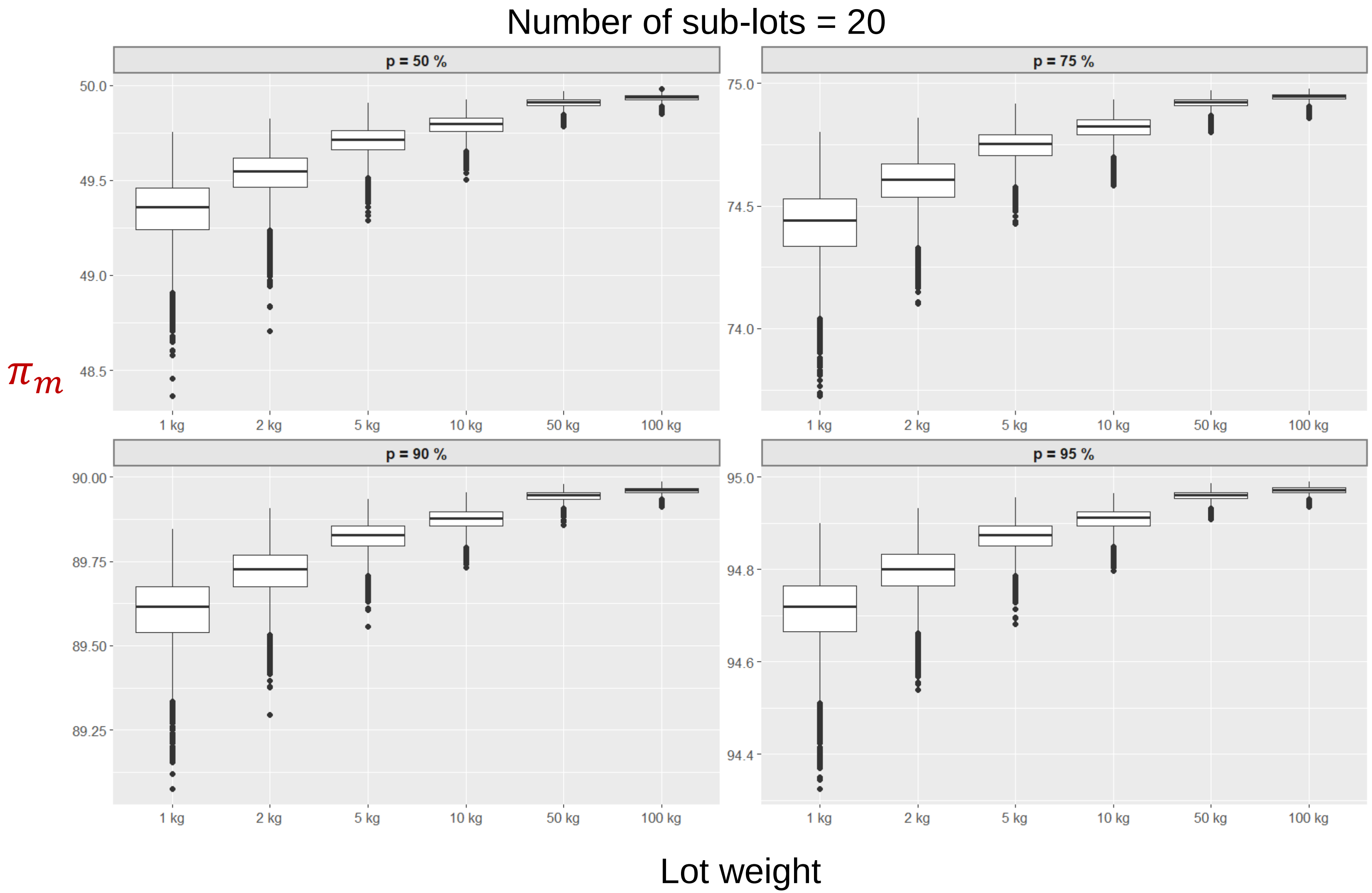
Sub-lots weight/Number of sub-lots ($M = \frac{Lot\ weight}{Sub-lot\ weight}$)

The larger the number of sub-lots M, the lower the minimum values π_m

2. Number of sub-lots determination – Some details

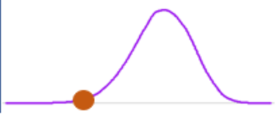
M true sub-lot values π_i from multivariate hypergeometric distribution

Minimum of the true M sub-lot values: $\pi_m = min\left(\{\pi_i\}_{i=1}^{M}\right)$



The lower the lot weight, the lower the minimum values π_m

Test value from another lab on the sub-lot with minimum value from a **Beta-binomial** distribution: p_2



Refresher:

- The result from a different lab is not from a **Binomial**(k, π_m) but from a distribution with a variance larger than the binomial variance
- The over-dispersion has been quantified by Miles (1963) and is taken into account in tolerance tables for comparing different laboratories

$$\sqrt{\frac{Over\_dispersed\_variance}{Binomial\_variance}} = f = 2.38 - 0.8321\pi_m$$

A model for generating over-dispersed binomial data is the Beta-binomial model with parameters k, α and β:

$$\alpha = \pi_m\left(\frac{k-1}{f^2-1} - 1\right) \qquad \beta = \alpha\left(\frac{1}{\pi_m} - 1\right)$$

Estimate of lot value: p from k seeds

Determine if p and p_2 are within tolerance according to ISTA Tolerance Table 5F

Generated from binomial distribution with parameters k and π

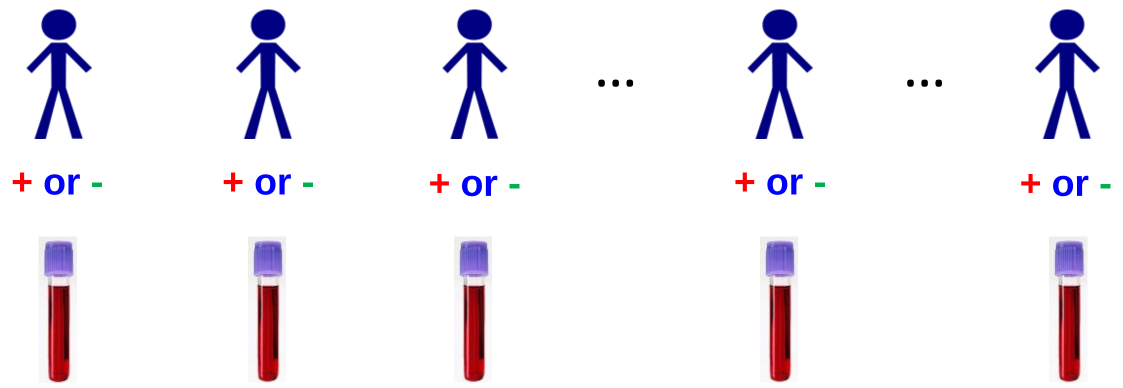
p and p_2 are independent

π (%)	W_{Lot} = 1.5 kg, W_{Sub} = 0.1 kg (M = 15)	W_{Lot} = 50 kg, W_{Sub} = 1 kg (M = 50)	W_{Lot} = 50 kg, W_{Sub} = 0.1 kg (M = 500)
	Prob(p and p_2 are within Tol)	Prob(p and p_2 are within Tol)	Prob(p and p_2 are within Tol)
50	0.9865	0.9857	0.9887
55	0.9858	0.9854	0.9836
60	0.9870	0.9871	0.9867
65	0.9852	0.9866	0.9854
70	0.9850	0.9853	0.9860
75	0.9840	0.9845	0.9842
80	0.9848	0.9858	0.9832
85	0.9861	0.9851	0.9849
90	0.9861	0.9892	0.9887
95	0.9893	0.9894	0.9885
99	0.9951	0.9959	0.9958

- For two extreme lot sizes (1.5 kg and 50 kg) and different number of sub-lots (15, 50 and 500), all the probabilities are very high (**above 0.98**)
- Evidence that given that the original lot is **homogeneous**, there is **no limit** in the number of sub-lots that can be elaborated from it

3. Group testing: number of groups to ensure that estimation is possible

- Suppose people are tested for a disease
- Who has the disease? → identification
  What is the prevalence of the disease? → estimation
  (i.e. what is the proportion of people with the disease?)
- One solution: individual testing



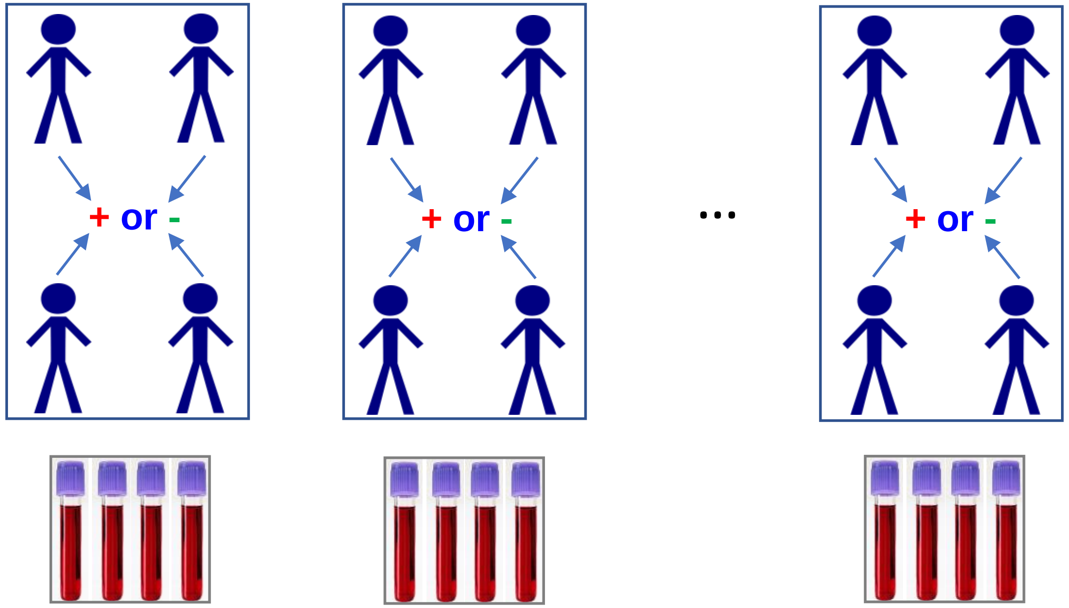
Identification: **population**

Estimation: **sample**

Analysis performed on blood samples

- Problem: can be expensive
- Group testing: cost savings

- Group testing:



Analysis performed on
mixed blood samples

- Identification: if the group is positive, the individuals making up the group are retested to determine which of the members have the disease.

- Estimation: $\hat{p}$

- Identification: original development of group testing by Robert Dorfman in 1943:

The Detection of Defective Members of Large Populations
Author(s): Robert Dorfman
Source: *The Annals of Mathematical Statistics*, Vol. 14, No. 4 (Dec., 1943), pp. 436-440

The inspection of the individual members of a large population is an expensive and tedious process. Often in testing the results of manufacture the work can be reduced greatly by examining only a sample of the population and rejecting the whole if the proportion of defectives in the sample is unduly large. In many inspections, however, the objective is to eliminate all the defective members of the population. This situation arises in manufacturing processes where the defect being tested for can result in disastrous failures. It also arises in certain inspections of human populations. Where the objective is to weed out individual defective units, a sample inspection will clearly not suffice. It will be shown in this paper that a different statistical approach can, under certain conditions, yield significant savings in effort and expense when a complete elimination of defective units is desired.

...

The method will be described by showing its application to a large-scale project on which the United States Public Health Service and the Selective Service System are now engaged. The object of the program is to weed out all syphilitic men called up for induction. Under this program each prospective inductee is subjected to a "Wasserman-type" blood test. The test may be divided conveniently into two parts:

...

- p : proportion of individuals in the population with the attribute
- n : number of groups
- m : number of individuals per group
- x : number of positive groups out of m

$$\hat{p} = 1 - \left(1 - \frac{x}{n}\right)^{\frac{1}{m}}$$

When all the groups are positive, $\hat{p} = 1$ and the result is not considered.

Y_i : random variable = 1 if group i is positive
= 0 if group i is negative

➡ Y_i has a Bernoulli distribution with pmf :

$$P(Y_i = y) = \begin{cases} (1-p)^m & \text{if } y = 0 \\ 1-(1-p)^m & \text{if } y = 1 \end{cases}$$

$P(\text{individual 1 neg \& individual 2 neg \& ... \& individual } m \text{ neg})$
$= P(\text{individual 1 neg}) \times P(\text{individual 2 neg}) \times \cdots \times P(\text{individual } m \text{ neg})$
$= (1-p) \times (1-p) \times \cdots \times (1-p)$

➡ $X = \sum_{i=1}^{n} Y_i$ has a binomial distribution $B(n, 1-(1-p)^m)$

Maximum Likelihood Estimation:

Functional invariance of MLEs property:
If $\hat{\theta}$ is the MLE for θ, and if $g(\theta)$ is any transformation of θ, then the MLE for $\alpha = g(\theta)$ is : $\hat{\alpha} = g(\hat{\theta})$.

Let $\theta = 1-(1-p)^m$.

Rearranging: $p = 1-(1-\theta)^{\frac{1}{m}} = g(\theta)$

The random variable X giving the number of positive groups x out of n has a binomial distribution $B(n, \theta)$.

Then : $\hat{\theta} = \frac{x}{n}$ is the MLE for θ .

➡ $\hat{p} = g(\hat{\theta}) = 1 - \left(1 - \frac{x}{n}\right)^{\frac{1}{m}}$

- Taking into account assay errors:

  λ : false negative (group tests neg when at least 1 individual is pos) rate = 1 – sensitivity

  δ : false positive (group tests pos when all individuals are neg) rate = 1 – specificity

$$\hat{p} = 1 - \left(1 - \frac{\frac{x}{n} - \delta}{1 - \lambda - \delta}\right)^{1/m}$$

with x being the number of groups out of n testing positive

See:

Statistical considerations in seed purity testing for transgenic traits

Seed Science Research (2001) 11, 101–119

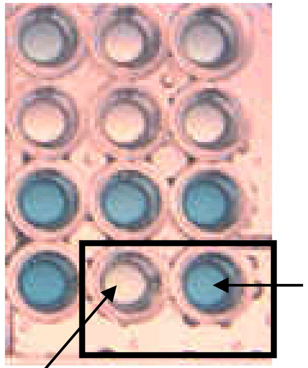
Kirk M. Remund[1]*, Doris A. Dixon[1], Deanne L. Wright[2] and Larry R. Holden[3]

Seed lot

10 groups of **150** seeds are taken from the lot using an appropriate sampling procedure (e.g. ISTA)

Seeds are ground into flour

Each sample is tested for **presence/absence** of GM seeds



6 groups are negative

4 groups are positive

6 groups of
150 ***seeds***
are negative

4 groups of
150 ***seeds***
are positive

Point estimate of *p*, the true proportion of GM seeds in the lot

$$\hat{p} = 1 - \left(1 - \frac{4}{10}\right)^{\frac{1}{150}}$$

$$= 0.34\%$$

- p : proportion of individuals in the population with the attribute
- n : number of groups
- m : number of individuals per group
- x : number of positive groups out of m

$$\hat{p} = 1 - \left(1 - \frac{x}{n}\right)^{\frac{1}{m}}$$

When all the groups are positive, estimation is not possible!

The probability that all groups are positive could help to ensure that the testing plan (n, m) will ensure estimation of p

Probability that all groups are positive: 2 cases

1. Infinite population size (e.g. seed lot)

2. Finite population size (e.g. sample distributed for a Proficiency Test)

1. Infinite population size: easy

k groups of m balls are sampled from a population of balls with a proportion π of white balls. The random variable Y_i "*number of white balls in group i*" has a binomial distribution with parameters m and π:

$$P(Y_i = n_i) = \binom{m}{n_i} \pi^{n_i} (1-\pi)^{m-n_i}\,.$$

Let A_i be the event "*the i^{th} group has at least one white ball*". Then, the probability that the 1st group is positive is:

$$P(A_1) = 1 - \binom{m}{0} \pi^0 (1-\pi)^m = 1-(1-\pi)^m\,.$$

The probability that the 1st and the 2nd groups have at least one white ball is:

$$P(A_1 \cap A_2) = (1-(1-\pi)^m)(1-(1-\pi)^m) = (1-(1-\pi)^m)^2.$$

The probability that all the groups have at least one white balls (i.e. that all the groups are positive) is:

$$P\left(\bigcap_{i=1}^{k} A_i\right) = (1-(1-\pi)^m)^k$$

2. Finite population size: less easy

n_1 white balls and n_2 black balls ($n_1 + n_2 = n$) are placed into k bins of maximum capacity m; $km = n$. Let X be the random variable "*number of bins without any white balls*". The random variable Y "*number of white balls in a sample of m balls*" has a hypergeometric distribution with parameters n, n_1 and m.

$$P(Y = w) = \frac{\binom{n_1}{w}\binom{km-n_1}{m-w}}{\binom{km}{m}}.$$

Let A_1 be the event "*the 1st sample has no white ball*". The probability that the 1st sample has no white ball is:

$$P(A_1) = \frac{\binom{km-n_1}{m}}{\binom{km}{m}} = \frac{\binom{km-m}{n_1}}{\binom{km}{n_1}}.$$

The probability that the 1st and the 2nd samples have no white ball is:

$$P(A_1 \cap A_2) = \frac{\binom{km-n_1}{m}}{\binom{km}{m}} \times \frac{\binom{km-n_1-m}{m}}{\binom{km-m}{m}} = \frac{\binom{km-2m}{n_1}}{\binom{km}{n_1}}.$$

The probability that the first s samples ($s < k$) have no white balls is:

$$P(\cap_{i=1}^{s} A_i) = \frac{\binom{km-sm}{n_1}}{\binom{km}{n_1}}.$$

The probability that any s particular bins have no white balls is:

$$P\left(\bigcap_{\substack{i\in I\\ I\subset\{1,2,\dots,k\}\\ \mathrm{Card}(I)=s}} A_i\right) = \binom{k}{s}\frac{\binom{km-sm}{n_1}}{\binom{km}{n_1}} = S_s$$

(there are $\binom{k}{s}$ possible combinations for s (out of k) bins without white balls).

Probability that at least one bin has no white ball:

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k}(-1)^{i+1}S_i \quad \text{(principle of inclusion-exclusion for probability)}$$

$$= \sum_{i=1}^{k}(-1)^{i+1}\binom{k}{i}\frac{\binom{km-im}{n_1}}{\binom{km}{n_1}}$$

$$= \frac{1}{\binom{km}{n_1}}\sum_{i=1}^{k}(-1)^{i+1}\binom{k}{i}\binom{m(k-i)}{n_1}.$$

The probability of having no bin without any white balls is:

$$P(X=0) = 1 - \frac{1}{\binom{km}{n_1}}\sum_{i=1}^{k}(-1)^{i+1}\binom{k}{i}\binom{m(k-i)}{n_1} = \frac{1}{\binom{km}{n_1}}\sum_{i=0}^{k}(-1)^{i}\binom{k}{i}\binom{m(k-i)}{n_1}$$

And therefore the probability that all the groups are positive is:

$$1 - \frac{1}{\binom{km}{n_1}}\sum_{i=0}^{k}(-1)^{i}\binom{k}{i}\binom{m(k-i)}{n_1}$$

An Excel calculator has been developed with an implementation of these computations as well as the computation of the expected number of positive groups:

	A	B
1	Group testing: on the number of positive groups	
2	Hypothesis: infinite population	
3		
4	Number of groups	10
5	Number of units per group	300
6	True characteristic content (%)	0.50%
7		
8	Probability that all groups are positive	8.09%
9	Expected number of positive groups	7.8
10		
11	Change any value in a yellow cell	
12		
13		
14		
15		
16		

Infinite population | Finite population | +

	A	B
1	Group testing: on the number of positive groups	
2	Hypothesis: finite population (size = 3000 units)	
3		
4	Number of groups	10
5	Number of units per group	300
6	True characteristic (%)	0.50%
7		
8	Probability that all groups are positive	4.66%
9	Expected number of positive groups	7.9
10		
11	Change any value in a yellow cell	
12		
13		
14		
15		
16		

Infinite population | Finite population | +

Not yet published

4. Opportunities

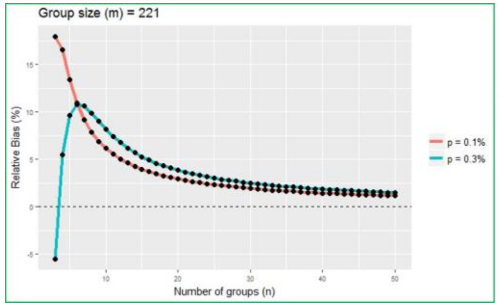
Revisiting group testing estimator properties: $\mathrm{Var}[\hat{p}] = \mathrm{E}[(\hat{p} - \mathrm{E}[\hat{p}])^2]$,

$\mathrm{Bias}[\hat{p}] = \mathrm{E}[\hat{p}] - p$,

$\mathrm{MSE}[\hat{p}] = \mathrm{E}[(\hat{p} - p)^2]$

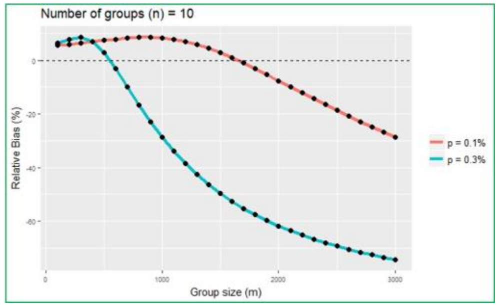
$$\mathrm{Bias}[\hat{p}] = \left[\sum_{x=0}^{n-1}\left(1-\left(1-\frac{x}{n}\right)^{1/m}\right)\binom{n}{x}\frac{(1-(1-p)^m)^x(1-p)^{m(n-x)}}{1-(1-(1-p)^m)^n}\right] - p$$

New insights for number of groups and group sizes recommendations



Recommendations on the number of groups to avoid under-estimation

Recommendations on group size to avoid under-estimation



Estimator of the proportion of the number of white balls from the observed number of empty bins for the hypergeometric group testing problem

Needs to solve the equation for n_1 :

$$d = k\left(1 - \frac{\binom{m(k-1)}{n_1}}{\binom{km}{n_1}}\right)$$

Difficult

where d is the observed number of positive bins.

Table 2.1. Minimum sampling intensity for seed lots in containers holding up to and including 100 kg seed

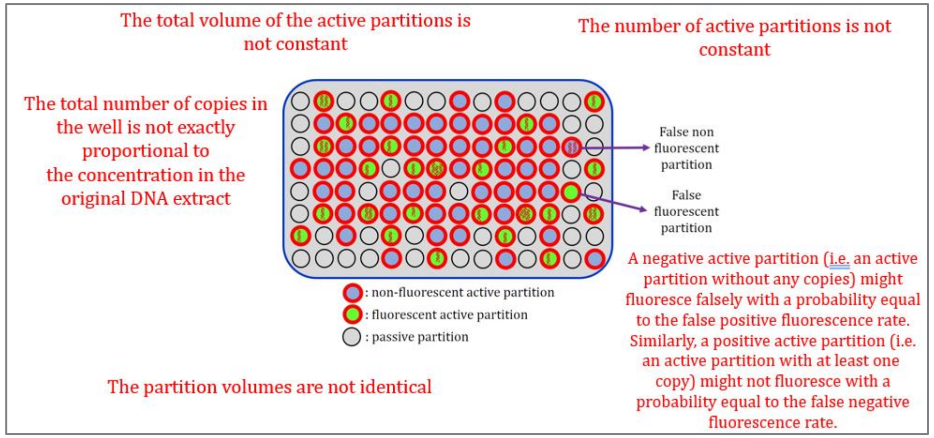
Number of containers	Minimum number of primary samples to be taken
1–4	3 primary samples from each container
5–8	2 primary samples from each container
9–15	1 primary sample from each container
16–30	15 primary samples, one each from 15 different containers
31–59	20 primary samples, one each from 20 different containers
60 or more	30 primary samples, one each from 30 different containers

These numbers have been elaborated over the years, using the results from sampling experiments and results from simulation studies.

Can we fine-tune these numbers using sampling theory?
(e.g. taking into account the size of the primary samples?)

Use of theoretical results on two-stage sampling?

- dPCR modeling



- Method validation:
  - Revising *ISTAgermMV* R package
  - Reviewing needs in terms of number of labs, number of lots,…
- …

4. Suggestions ?

Thank you!

Follow us on social media: