Statistical modelling in ISTA

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

… all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind….

George Box, 1987

Outcomes from a statistical model: estimates

Point estimate: e.g. mean

Variance of the estimate

$\sqrt{\text{Variance of the estimate}}$ → standard uncertainty

$$\text{Main statistical model used in ISTA: linear } \left\{\begin{array}{c}\text{fixed}\\ \text{random}\\ \text{mixed}\end{array}\right\} \text{ effects model}$$

(Example: **AN**alysis **O**f **VA**riance model: linear fixed effects model)

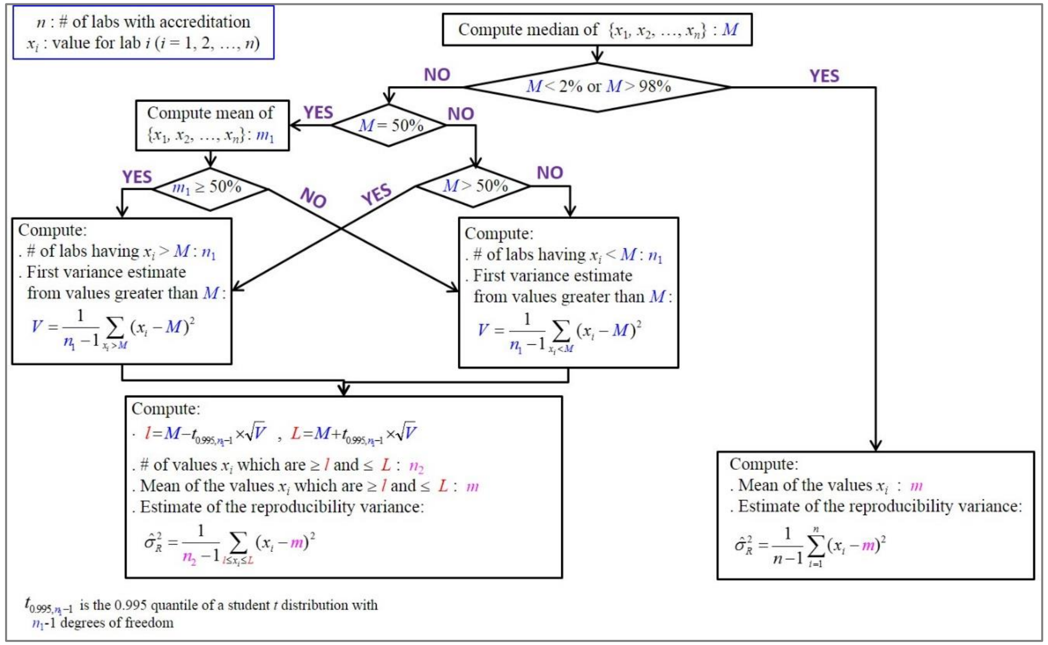
Examples of use of the linear $\left\{\begin{array}{c}\text{fixed}\\ \text{random}\\ \text{mixed}\end{array}\right\}$ effects model in ISTA

1. **Germination Proficiency Tests**:

Ratings are based on z-scores:

$$z_i = \frac{x_i - m}{\hat{\sigma}_R}$$

The denominator of the z-scores can be viewed as the **reproducibility** standard deviation estimated from a linear random effects model:



n : # of labs with accreditation
x_i : value for lab i ($i = 1, 2, \ldots, n$)

Compute median of $\{x_1, x_2, \ldots, x_n\}$: M

$M < 2\%$ or $M > 98\%$

$M = 50\%$

Compute mean of $\{x_1, x_2, \ldots, x_n\}$: m_1

$m_1 \geq 50\%$

$M > 50\%$

Compute:
. # of labs having $x_i > M$: n_1
. First variance estimate from values greater than M :

$$V = \frac{1}{n_1 - 1}\sum_{x_i > M}(x_i - M)^2$$

Compute:
. # of labs having $x_i < M$: n_1
. First variance estimate from values greater than M :

$$V = \frac{1}{n_1 - 1}\sum_{x_i < M}(x_i - M)^2$$

Compute:
. $l = M - t_{0.995, n_1 - 1} \times \sqrt{V}$, $L = M + t_{0.995, n_1 - 1} \times \sqrt{V}$
. # of values x_i which are $\geq l$ and $\leq$ L : n_2
. Mean of the values x_i which are $\geq l$ and $\leq$ L : m
. Estimate of the reproducibility variance:

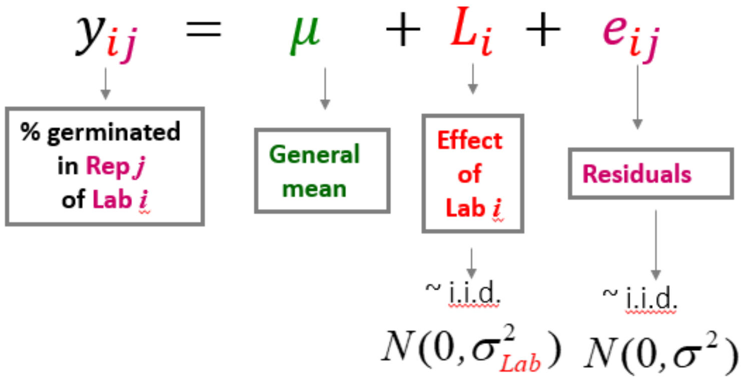
$$\hat{\sigma}_R^2 = \frac{1}{n_2 - 1}\sum_{l \leq x_i \leq L}(x_i - m)^2$$

Compute:
. Mean of the values x_i : m
. Estimate of the reproducibility variance:

$$\hat{\sigma}_R^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$$

$t_{0.995, n_1 - 1}$ is the 0.995 quantile of a student t distribution with n_1-1 degrees of freedom

NO
YES
YES
NO
YES
NO
NO
YES



$$y_{ij} = \mu + L_i + e_{ij}$$

% germinated in Rep j of Lab i
General mean
Effect of Lab i
Residuals
~ i.i.d. $N(0, \sigma^2_{Lab})$
~ i.i.d. $N(0, \sigma^2)$

Consistent with ISO 5725-2, 1994

→ **Reproducibility** std-dev: $\widehat{\sigma_R} = \sqrt{\widehat{\sigma^2_{Lab}} + \widehat{\sigma^2}/\#\_of\_reps}$

Examples of use of the linear { fixed / random / mixed } effects model in ISTA

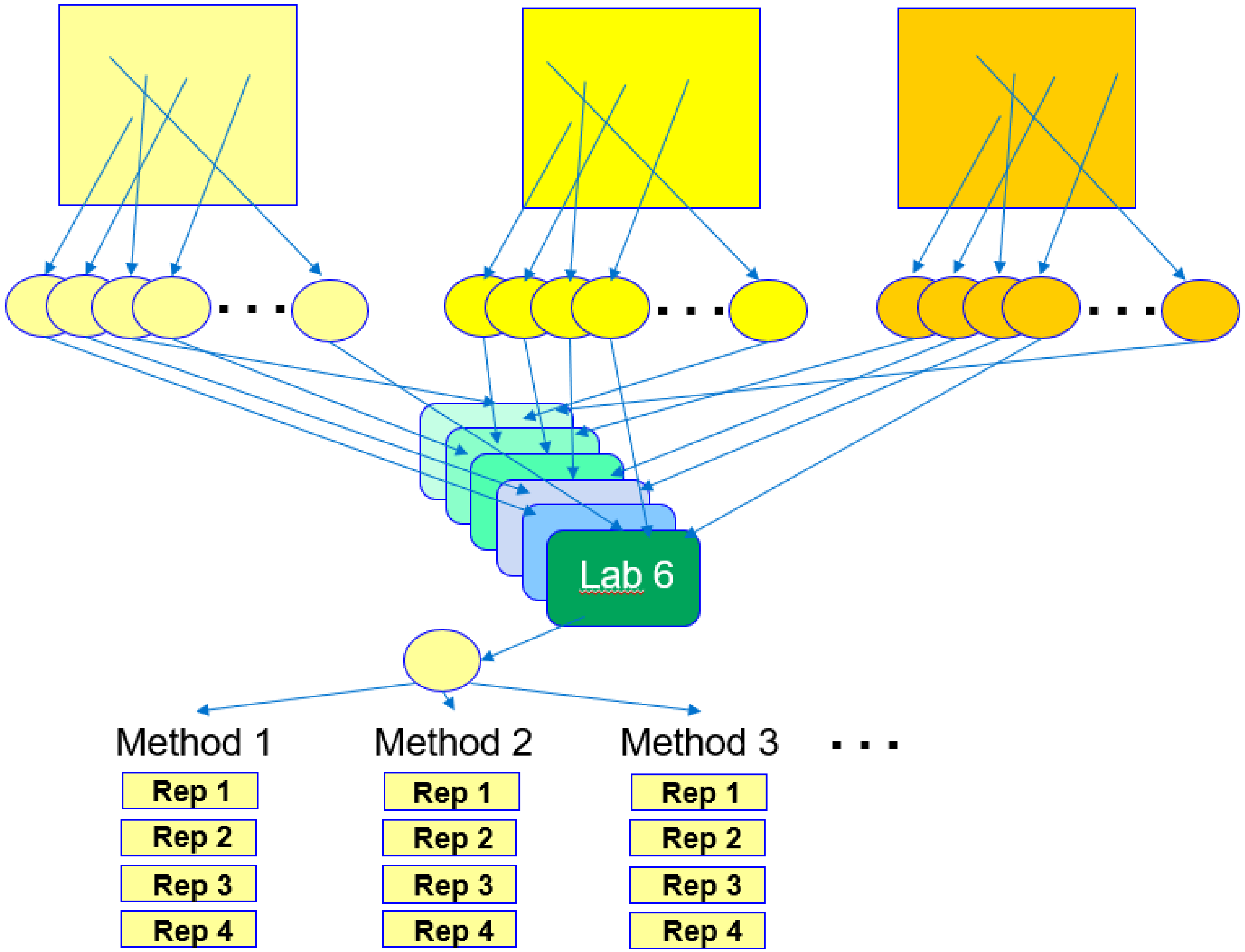
2. Analysis of germination Method Validation studies:

Seed lots with different quality levels (usually **3**)

Samples of seeds are elaborated

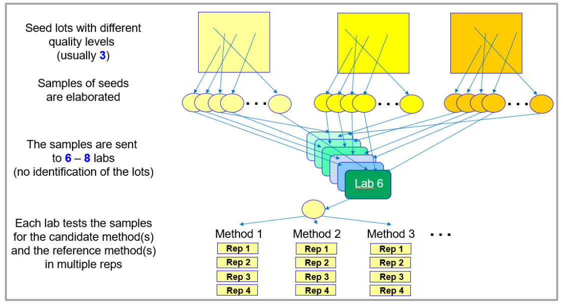
The samples are sent to **6** – **8** labs (no identification of the lots)

Each lab tests the samples for the candidate method(s) and the reference method(s) in multiple reps



Examples of use of the linear $\left\{\begin{matrix}\text{fixed}\\ \text{random}\\ \text{mixed}\end{matrix}\right\}$ effects model in ISTA

2. Analysis of germination Method Validation studies:



Seed lots with different quality levels (usually **3**)

Samples of seeds are elaborated

The samples are sent to **6** – **8** labs (no identification of the lots)

Each lab tests the samples for the candidate method(s) and the reference method(s) in multiple reps

Assessing **repeatability**/**reproducibility** for each method:

→ For each method, fit the following linear mixed effects model:

$$y_{ijk} = \pi + \alpha_i + L_j + (\alpha L)_{ij} + e_{ijk}$$

y_{ijk} → **% germinated in Rep *k* of Lot *i* and Lab *j***

π → **General mean**

α_i → **Effect of Lot *i***

L_j → **Effect of Lab *j*** ~ i.i.d. $N(0, \sigma^2_{Lab})$

$(\alpha L)_{ij}$ → **Interaction effect between Lot *i* and Lab *j*** ~ i.i.d. $N(0, \sigma^2_{Lot \times Lab})$

e_{ijk} → **Residuals** ~ i.i.d. $N(0, \sigma^2)$

→ **Repeatability** std-dev: $\sqrt{\widehat{\sigma^2}}$

Model for 1 lab: $y_{ik} = \pi + \alpha_i + e_{ik}$

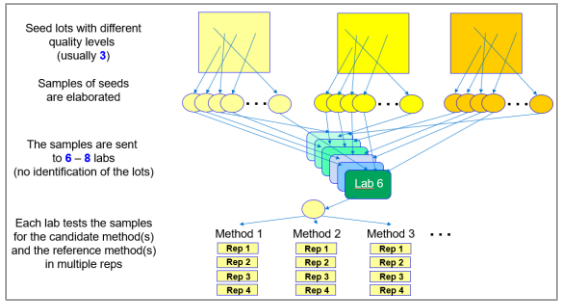
then: $\mathrm{Var}[y_{ik}] = \mathrm{Var}[\pi + \alpha_i + e_{ik}] = \mathrm{Var}[e_{ik}] \approx \widehat{\sigma^2}$

→ **Reproducibility** std-dev:

of the **mean** over K reps: $\sqrt{\widehat{\sigma^2_{Lab}} + \widehat{\sigma^2_{Lot \times Lab}} + \widehat{\sigma^2}/K}$

$$\begin{aligned}\mathrm{Var}[\overline{y_{ij.}}] &= \mathrm{Var}\left[\pi + \alpha_i + L_j + (\alpha L)_{ij} + \frac{1}{K}\sum_{k=1}^{K} e_{ijk}\right]\\ &= \mathrm{Var}[L_j] + \mathrm{Var}[(\alpha L)_{ij}] + \tfrac{1}{K^2}\textstyle\sum_{k=1}^{K}\mathrm{Var}[e_{ijk}]\\ &\approx \widehat{\sigma^2_{Lab}} + \widehat{\sigma^2_{Lot \times Lab}} + \widehat{\sigma^2}/K\end{aligned}$$

Examples of use of the linear $\left\{\begin{matrix}\text{fixed}\\ \text{random}\\ \text{mixed}\end{matrix}\right\}$ effects model in ISTA



Seed lots with different quality levels (usually 3)

Samples of seeds are elaborated

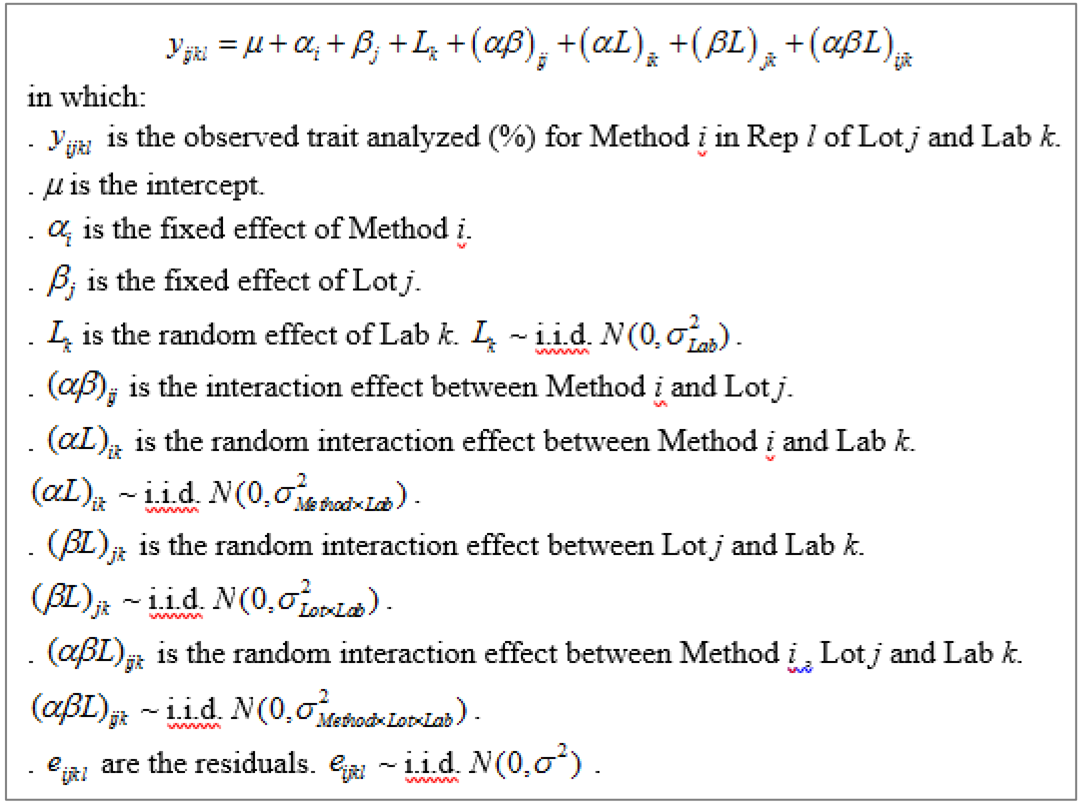
The samples are sent to 6 – 8 labs (no identification of the lots)

Each lab tests the samples for the candidate method(s) and the reference method(s) in multiple reps

2. Analysis of germination Method Validation studies:

Comparing Method, Lot, and Method x Lot means:

Fit the linear mixed effects model:



$$y_{ijkl} = \mu + \alpha_i + \beta_j + L_k + (\alpha\beta)_{ij} + (\alpha L)_{ik} + (\beta L)_{jk} + (\alpha\beta L)_{ijk}$$

in which:

. y_{ijkl} is the observed trait analyzed (%) for Method i in Rep l of Lot j and Lab k.

. μ is the intercept.

. α_i is the fixed effect of Method i.

. β_j is the fixed effect of Lot j.

. L_k is the random effect of Lab k. $L_k \sim$ i.i.d. $N(0, \sigma^2_{Lab})$.

. $(\alpha\beta)_{ij}$ is the interaction effect between Method i and Lot j.

. $(\alpha L)_{ik}$ is the random interaction effect between Method i and Lab k.

$(\alpha L)_{ik} \sim$ i.i.d. $N(0, \sigma^2_{Method\times Lab})$.

. $(\beta L)_{jk}$ is the random interaction effect between Lot j and Lab k.

$(\beta L)_{jk} \sim$ i.i.d. $N(0, \sigma^2_{Lot\times Lab})$.

. $(\alpha\beta L)_{ijk}$ is the random interaction effect between Method i, Lot j and Lab k.

$(\alpha\beta L)_{ijk} \sim$ i.i.d. $N(0, \sigma^2_{Method\times Lot\times Lab})$.

. e_{ijkl} are the residuals. $e_{ijkl} \sim$ i.i.d. $N(0, \sigma^2)$.

→ **ANOVA** table for the **fixed effects** (**Method**, **Lot** and **Method x Lot**)

→ **Least Squares (LS) Means** comparisons

Examples of use of the linear $\left\{\begin{array}{c}\text{fixed}\\ \text{random}\\ \text{mixed}\end{array}\right\}$ effects model in ISTA

2. Analysis of germination Method Validation studies:

Given that germination percentages have a binomial distribution, one could ask why are we not using **G**eneralized **L**inear **M**ixed effects **M**odel (**GLMM**) for the analysis?

→ Output from different GLMM algorithms (e.g. the ones implemented in SAS GLIMMIX procedure, in glmmPQL() from MASS R package, glmer() from lme4 R package, …) have been compared: they provide estimates that can be very different

→ When fitting a GLMMM, it is assumed that the random effects on the linear predictor scale are normally distributed: interpretation is not obvious as well as the transformation back to the data scale

→ Literature review : ISO organization has no specific recommendation,
few approaches found are not convincing

Best approach is to use a **L**inear **M**ixed effects **M**odel (**LMM**)

Other statistical models based on probability distributions

1. Modeling over-dispersion for non-commercial seed lots (e.g. wild species)

True proportion of germinated seeds in the lot: π

X_i = 1 if seed i germinates, 0 otherwise

Bernoulli variable

Sample of n seeds

↓

$Y = \sum_{i=1}^{n} X_i$: number of germinated seeds

Miles's dispersion factor : $f = \frac{\sigma}{\sigma_B}$

where $\sigma_B^2 = n\pi\,(1-\pi)$ and σ^2 is the variance among the reps of a germination test

Commercial seed lot

π is a constant

↓

Y ~ Binomial(n,π) with π being the true germination proportion in the seed lot

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.

↓

Mean(f) = 1

Median(f) ≤ 0.9

Non-commercial seed lot

$\pi \sim$ Beta(a,b)

(the probability for an individual seed to germinate is unknown or random)

↓

Y ~ Beta-Binomial(n, α, β) with

$\alpha = \pi\left(\frac{n-1}{f^2-1} - 1\right)$ and $\beta = \alpha\left(\frac{1}{\pi} - 1\right)$

↓

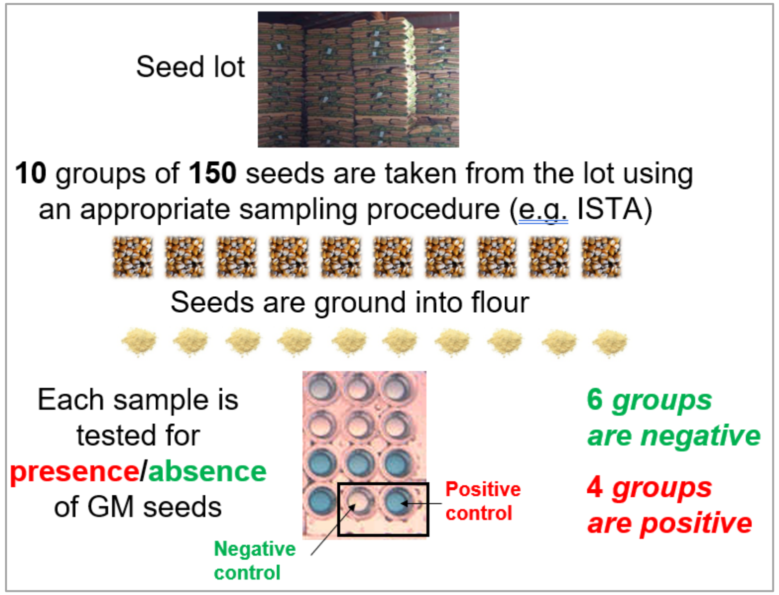
Over dispersed rep results:

$$\sigma^2 = n\pi(1-\pi)\left(1 + \frac{n-1}{\alpha+\beta+1}\right) = n\pi(1-\pi)f^2$$

Mean(f) >> 1

Other statistical models based on probability distributions

2. Group testing estimator



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

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

$$= 0.34\%$$

Derived from the distribution of the number of positive groups:

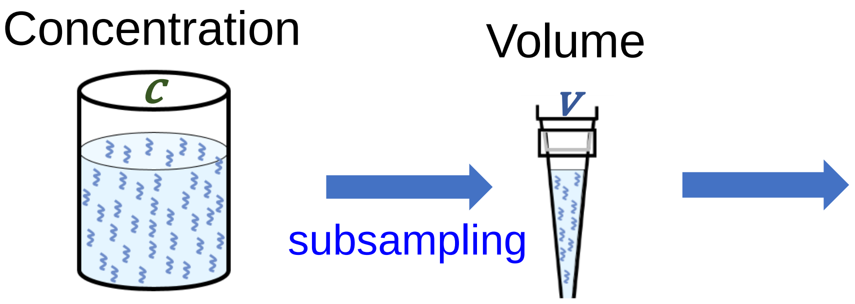
binomial distribution $B(n, 1-(1-p)^m)$

n : number of groups

m : number of individuals per group

Other statistical models based on probability distributions

3. Volume subsampling

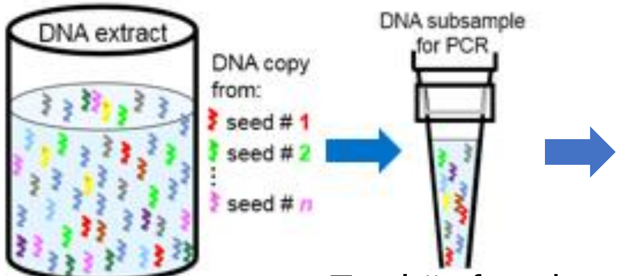


number of ⌇ in the subsample is **Poisson**(cv)

Siméon Poisson

Siméon Denis Poisson (1781–1840)

Max group size



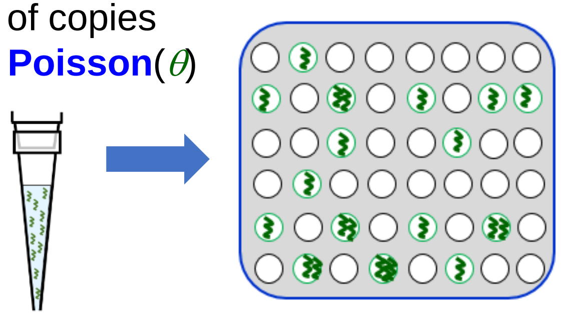
Total # of copies $T \sim$ **Poisson**(λ)

number of copies from each seed given a total $T = t$ of copies is **Multinomial**(t; $1/n, \ldots, 1/n$)

Poissonization theorem for multinomials: **unconditionally**, number of copies X_i for seed i is **Poisson**(λ/n) and X_i are independent

dPCR

# of copies $M \sim$ **Poisson**(θ)



number X of empty partitions given $M = m$ has a distribution with pmf:

$$P(X = x|M = m) = \binom{k}{x}\sum_{i=0}^{k-x}(-1)^i\binom{k-x}{i}\left(1-\frac{x+i}{k}\right)^m$$

Unconditionally, X is **Binomial**($k, e^{-\theta/k}$)

Simpler distributions!

Thank you!

INTERNATIONAL-SEED TESTING
UNIFORMITY IN SEED TESTING

Follow us on social media: