

# Supplément au chapitre 4

## Liste des statistiques descriptives

Gurvan Hermange

### Contenu de l'annexe

Cette annexe présente à la section 1 l'expression des différentes statistiques descriptives en fonction des variables d'observation.

Les tableaux 1 et 2 décrivent en quelques mots à quoi correspond chaque statistique descriptive.

La section 2 présente l'ensemble des figures où l'on montre les valeurs prises par nos statistiques descriptives.

### Table des matières

<b>1</b>	<b>Liste des statistiques descriptives</b>	<b>2</b>
1.1	Incucyte . . . . .	2
1.2	MultiGen . . . . .	5
<b>2</b>	<b>Valeurs prises par les statistiques descriptives</b>	<b>10</b>

# 1 Liste des statistiques descriptives

## 1.1 Incucyte

Pour un type  $p \in \{1, 2, 3\}$  de départ, les statistiques descriptives calculées à partir des variables d'observation associées à l'expérience Incucyte (et présentées dans le tableau 1) sont les suivantes :

$$Y_1^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} 1_{D_{1,i} < T_{max}}$$

$Y_1^p$  correspond à la proportion d'observations des temps de première division non censurées à droite, à cause de la limite du temps d'observation. On fait de même pour les observations des temps de secondes divisions.

$$Y_2^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} 1_{D_{2,1,i} < T_{max}}$$

$$Y_3^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} 1_{D_{2,2,i} < T_{max}}$$

$$Y_4^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} D_{2,1,i}$$

$$Y_5^p = \sqrt{\frac{1}{N_{\mathcal{I}_p} - 1} \sum_{i \in \mathcal{I}_p} (D_{2,1,i} - Y_4^p)^2}$$

qui correspond à l'écart type empirique.

$$Y_6^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} \left( \frac{D_{2,1,i} - Y_4^p}{Y_5^p} \right)^3$$

qui correspond au "skewness".

$$Y_7^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} \left( \frac{D_{2,1,i} - Y_4^p}{Y_5^p} \right)^4$$

qui correspond au "kurtosis". On fait de même en définissant  $Y_8^p, Y_9^p, Y_{10}^p$  et  $Y_{11}^p$  pour  $D_{2,2}$ . Puis :

$$Y_{12}^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} \frac{1}{2} (D_{2,1,i} + D_{2,2,i})$$

et  $Y_{13}^p, Y_{14}^p$  et  $Y_{15}^p$  qui vont respectivement correspondre à l'écart-type empirique, le skewness empirique et le kurtosis empirique pour  $\frac{D_{2,1} + D_{2,2}}{2}$ .

On définit ensuite  $Y_{16}^p$  qui correspond à la corrélation empirique entre les deux échantillons  $(D_{2,2,i})_{i \in \mathcal{I}_p}$  et  $(D_{2,1,i})_{i \in \mathcal{I}_p}$ .  $Y_{17}^p$  correspond à la moyenne empirique de  $(D_{2,1,i} - D_{1,i})_{i \in \mathcal{I}_p}$  :

$$Y_{17}^p = \frac{1}{N_{\mathcal{I}_p}} \sum_{i \in \mathcal{I}_p} (D_{2,1,i} - D_{1,i})$$

$Y_{18}^p$  correspond à l'écart type empirique de  $(D_{2,1,i} - D_{1,i})_{i \in \mathcal{I}_p}$  :

$$Y_{18}^p = \sqrt{\frac{1}{N_{\mathcal{I}_p} - 1} \sum_{i \in \mathcal{I}_p} (D_{2,1,i} - Y_{17}^p)^2}$$

$Y_{19}^p$  et  $Y_{20}^p$  sont respectivement les moyenne et écart-type empirique de  $(D_{2,2,i} - D_{1,i})_{i \in \mathcal{I}_p}$ .  $Y_{21}^p$  correspond à la corrélation empirique entre les deux échantillons  $(D_{2,2,i} - D_{1,i})_{i \in \mathcal{I}_p}$  et  $(D_{2,1,i} - D_{1,i})_{i \in \mathcal{I}_p}$ .

$Y_{21}^p$  et  $Y_{22}^p$  sont respectivement la moyenne et l'écart-type empirique du nombre de cellules observées à  $T_{max}$  par l'expérience Incucyte (c'est-à-dire sans bruit d'échantillonnage) :

$$Y_{21}^p = \frac{1}{N_{\mathcal{I}}} \sum_{i \in \mathcal{I}} \text{card}(\mathcal{C}_{T_{max}, i})$$

$i$	label
1	prop_T1_uncensored
2	prop_T2_1_uncensored
3	prop_T2_2_uncensored
4	mean_T2_1
5	std_T2_1
6	skewness_T2_1
7	kurtosis_T2_1
8	mean_T2_2
9	std_T2_2
10	skewness_T2_2
11	kurtosis_T2_2
12	mean_T2
13	std_T2
14	skewness_T2
15	kurtosis_T2
16	cor_T2
17	mean_T2_1_minus_T1
18	std_T2_1_minus_T1
19	mean_T2_2_minus_T1
20	std_T2_2_minus_T1
21	cor_T2_minus_T1
22	mean_colony_size
23	std_colony_size

TABLE 1 – Liste de toutes les statistiques descriptives considérées initialement et étant associées aux observations issues de l'expérience Incucyte :  $Y_i$  (on ne mentionne pas ici le type de la cellule de départ) pour  $i \in \{1, \dots, 23\}$  (première colonne) et le label associé qui décrit succinctement la façon dont la statistique descriptive est calculée. Le label et l'index de la statistique descriptive sont ceux qui seront indiqués sur les figures.

## 1.2 MultiGen

Pour un type  $p \in \{1, 2\}$  de départ, nous présentons ci-dessous les statistiques descriptives calculées à partir des variables d'observation associées à l'expérience MultiGen (voir Tab. 2) et utilisées dans ce travail.

Notons que nous n'utilisons pas ici de données correspondant à des expériences MultiGen faites avec des HPC comme cellules de départ.

$Y_{24}^p$  correspond au nombre moyen de cellules à avoir subi zéro division, sur les cellules échantillonnées correspondant à celles observées dans l'expérience MultiGen :

$$Y_{24}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=0}$$

On définit de même  $Y_{25}^p, Y_{26}^p, Y_{27}^p, Y_{28}^p$  et  $Y_{29}^p$  pour des nombres de divisions supérieurs :

$$Y_{25}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=1}$$

$$Y_{26}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=2}$$

$$Y_{27}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=3}$$

$$Y_{28}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=4}$$

$$Y_{29}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=5}$$

sachant qu'on ne considère pas de divisions observables au delà de 5. Ainsi, si une cellule se divise plus, elle sera comptée comme ayant subi un nombre de 5 divisions.

On définit les écart-types associés :

$$Y_{30}^p = \sqrt{\frac{1}{N_{\mathcal{J}_p} - 1} \sum_{j \in \mathcal{J}} \left[ \left( \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=0} \right) - Y_{24}^p \right]^2}$$

Et de même pour  $Y_{31}^p, Y_{32}^p, Y_{33}^p, Y_{34}^p$  et  $Y_{36}^p$ .

On définit  $Y_{36}^p$  comme étant la proportion de familles dont toutes les cellules observées n'ont fait aucune division :

$$Y_{36}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=0}$$

On définit de même  $Y_{37}^p$  comme étant la proportion de familles dont toutes les cellules observées ont fait une et une seule division :

$$Y_{37}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=1}$$

On définit de même  $Y_{38}^p, Y_{39}^p, Y_{40}^p$  et  $Y_{41}^p$  pour des nombres de divisions supérieurs :

$i$	label
24	mean_nb_cells_gen1
25	mean_nb_cells_gen2
26	mean_nb_cells_gen3
27	mean_nb_cells_gen4
28	mean_nb_cells_gen5
29	mean_nb_cells_gen6
30	std_nb_cells_gen1
31	std_nb_cells_gen2
32	std_nb_cells_gen3
33	std_nb_cells_gen4
34	std_nb_cells_gen5
35	std_nb_cells_gen6
36	prop_fam_only_gen_1
37	prop_fam_only_gen_2
38	prop_fam_only_gen_3
39	prop_fam_only_gen_4
40	prop_fam_only_gen_5
41	prop_fam_only_gen_6
42	prop_fam_rep_over_6gen
43	prop_fam_rep_over_5gen
44	prop_fam_rep_over_4gen
45	prop_fam_rep_over_3gen
46	prop_fam_rep_over_2gen
47	prop_fam_rep_over_1gen
48	mean_nb_cells_norm
49	std_nb_cells_norm
50	mean_nb_cells_type_HSC
51	mean_nb_cells_type_MPP
52	mean_nb_cells_type_HPC
53	mean_nb_cells_type_CD34neg
54	std_nb_cells_type_HSC
55	std_nb_cells_type_MPP
56	std_nb_cells_type_HPC
57	std_nb_cells_type_CD34neg
58	prop_fam_only_HSC
59	prop_fam_only_MPP
60	prop_fam_only_HPC
61	prop_fam_only_CD34neg
62	prop_fam_rep_over_4types
63	prop_fam_rep_over_3types
64	prop_fam_rep_over_2types
65	prop_fam_rep_over_1types
66	mean_nb_cells
67	std_nb_cells

TABLE 2 – Liste de toutes les statistiques descriptives considérées initialement et étant associées aux observations issues de l'expérience MultiGen :  $Y_i$  (on ne mentionne pas ici le type de la cellule de départ) pour  $i \in \{24, \dots, 67\}$  (première colonne) et le label associé qui décrit succinctement la façon dont la statistique descriptive est calculée. Le label et l'index de la statistique descriptive sont ceux qui seront indiqués sur les figures.

$$Y_{38}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=2}$$

$$Y_{39}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=3}$$

$$Y_{40}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=4}$$

$$Y_{41}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{n_k=5}$$

On définit  $Y_{42}^p$  comme étant la proportion de familles dont les cellules observées sont réparties sur exactement 6 générations, c'est-à-dire qu'on observe au moins 6 cellules ayant chacune subi un nombre de division différent. Pour cela, notons, pour une famille  $j \in \mathcal{J}$ , le vecteur  $X_j$  de taille 6 tel que, pour  $n \in \{0, \dots, 5\}$  :

$$X_{j,n} = \begin{cases} 1 & \text{si } \exists k \in \hat{\mathcal{C}}_{T_{max},j}, n_k = n \\ 0 & \text{sinon} \end{cases}$$

Alors :

$$Y_{42}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n X_{j,n}=6}$$

On définit de même  $Y_{43}^p$  comme étant la proportion de familles dont les cellules observées sont réparties sur exactement 5 générations :

$$Y_{43}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n X_{j,n}=5}$$

On définit  $Y_{44}^p$  comme étant la proportion de familles dont les cellules observées sont réparties sur exactement 4 générations :

$$Y_{44}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n X_{j,n}=4}$$

On définit  $Y_{45}^p$  comme étant la proportion de familles dont les cellules observées sont réparties sur exactement 3 générations :

$$Y_{45}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n X_{j,n}=3}$$

On définit  $Y_{46}^p$  comme étant la proportion de familles dont les cellules observées sont réparties sur exactement 2 générations :

$$Y_{46}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n X_{j,n}=2}$$

On définit  $Y_{47}^p$  comme étant la proportion de familles dont les cellules observées sont réparties sur exactement 1 génération :

$$Y_{47}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n X_{j,n}=1}$$

Notons que  $Y_{47}^p = \sum_{36 \leq i \leq 41} Y_i^p$ .

$Y_{48}^p$  va correspondre à un nombre moyen "effectif" de cellules, c'est-à-dire en pondérant par le nombre de divisions (plus précisément, le nombre de générations, qui est égal au nombre de divisions + 1) :

$$Y_{48}^p = \sum_{1 \leq i \leq 6} \frac{Y_{24+i-1}^p}{i}$$

$Y_{49}^p$  correspond à l'écart type associé à  $Y_{48}^p$ .

On définit  $Y_{50}^p$  comme le nombre moyen de cellules (observées) à être de type HSC\* (partant d'une cellule initiale de type  $p$ ) :

$$Y_{50}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{p_k=1}$$

On définit  $Y_{51}^p$  comme le nombre moyen de cellules (observées) à être de type MPP :

$$Y_{51}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{p_k=2}$$

On définit  $Y_{52}^p$  comme le nombre moyen de cellules (observées) à être de type HPC :

$$Y_{52}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{p_k=3}$$

On définit  $Y_{53}^p$  comme le nombre moyen de cellules (observées) à être de type CD34<sup>-</sup> :

$$Y_{53}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{p_k=4}$$

On définit les écart-types associés :

$$Y_{54}^p = \sqrt{\frac{1}{N_{\mathcal{J}} - 1} \sum_{j \in \mathcal{J}} \left[ \left( \sum_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{p_k=1} \right) - Y_{50}^p \right]^2}$$

Et de même pour  $Y_{55}^p$ ,  $Y_{56}^p$  et  $Y_{57}^p$ .

On définit  $Y_{58}^p$  comme la proportion de familles à n'avoir que des cellules de type HSC\* (parmi celles observées) :

$$Y_{58}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{p_k=1}$$

On définit  $Y_{59}^p$  comme la proportion de familles à n'avoir que des cellules de type MPP (parmi celles observées) :

$$Y_{59}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{p_k=2}$$

On définit  $Y_{60}^p$  comme la proportion de familles à n'avoir que des cellules de type HPC (parmi celles observées) :

$$Y_{60}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{p_k=3}$$

On définit  $Y_{61}^p$  comme la proportion de familles à n'avoir que des cellules de type CD34<sup>-</sup> (parmi celles observées) :

$$Y_{61}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \prod_{k \in \hat{\mathcal{C}}_{T_{max},j}} 1_{p_k=4}$$

On définit  $Y_{62}^p$  comme étant la proportion de familles dont les cellules observées sont réparties sur exactement les 4 types cellulaires, c'est-à-dire qu'on observe au moins 4 cellules ayant chacune



subi un type cellulaire différent. Pour cela, notons, pour une famille  $j \in \mathcal{J}$ , le vecteur  $Z_j$  de taille 4 tel que, pour  $a \in \{1, 2, 3, 4\}$  :

$$Z_{j,a} = \begin{cases} 1 & \text{si } \exists k \in \hat{\mathcal{C}}_{T_{max,j}}, p_k = a \\ 0 & \text{sinon} \end{cases}$$

Alors :

$$Y_{62}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n Z_{j,n}=4}$$

On définit de même  $Y_{63}^p$ ,  $Y_{64}^p$ , et  $Y_{65}^p$  comme étant la proportion de familles dont les cellules observées sont respectivement réparties sur exactement 3, 2 et 1 génération(s) :

$$Y_{63}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n Z_{j,n}=3}$$

$$Y_{64}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n Z_{j,n}=2}$$

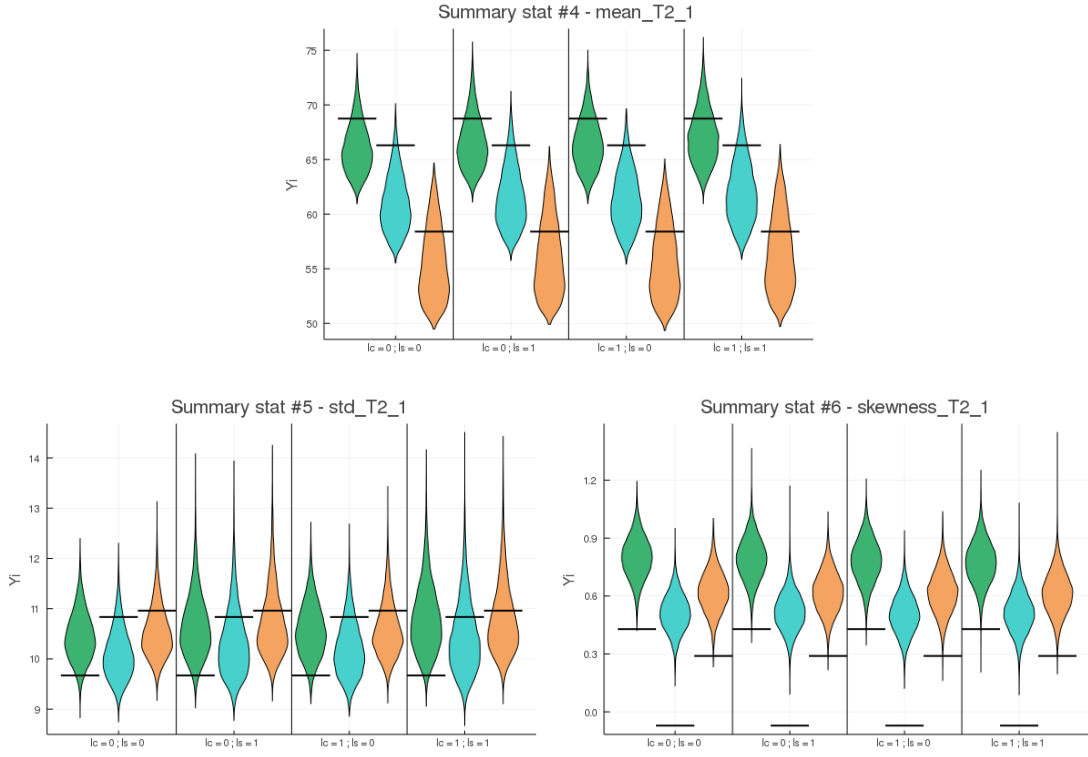
$$Y_{65}^p = \frac{1}{N_{\mathcal{J}}} \sum_{j \in \mathcal{J}} 1_{\sum_n Z_{j,n}=1}$$

Notons que nous avons  $Y_{65}^p = Y_{58}^p + Y_{59}^p + Y_{60}^p + Y_{61}^p$

Enfin, on définit  $Y_{66}^p$  et  $Y_{67}^p$  comme étant respectivement la moyenne et l'écart-type empirique du nombre de cellules observées par l'expérience MultiGen (donc avec bruit d'échantillonnage).

Notons que nous avons la relation :

$$\begin{aligned} Y_{66}^p &= Y_{50}^p + Y_{51}^p + Y_{52}^p + Y_{53}^p \\ &= Y_{24}^p + Y_{25}^p + Y_{26}^p + Y_{27}^p + Y_{28}^p + Y_{29}^p \end{aligned}$$



## 2 Valeurs prises par les statistiques descriptives

Nous présentons dans cette section les figures montrant les valeurs prises par les statistiques descriptives  $Y_i$  sur l'ensemble des paramètres échantillonnés par Latin Hypercube, suivant l'un des 4 modèles  $\mathcal{M}$  en abscisse (i.e. suivant l'une des 4 valeurs prises par le couple  $(I_c, I_s)$  modélisant la présence (valeur égale à 1) ou absence (0) de concordance et synchronicité) et l'une des 3 conditions initiales, à savoir la cellule de départ est de type HSC\* (vert), MPP (cyan) ou HPC (orange). Les lignes horizontales (représentées à l'identique pour chaque modèle) matérialisent la valeur de ces statistiques descriptives  $\hat{y}_i^p$  calculées sur les données expérimentales en fonction du type  $p$  de la cellule de départ.

Notons que nous excluons les figures correspondant aux statistiques descriptives 1, 2, 3, 24, 30, 42 et 43 pour lesquelles les visualisations en "violin plot" n'ont pas d'intérêt car les statistiques descriptives prennent soient des valeurs toutes (ou presque) égales à zéro ou égales à 1.

