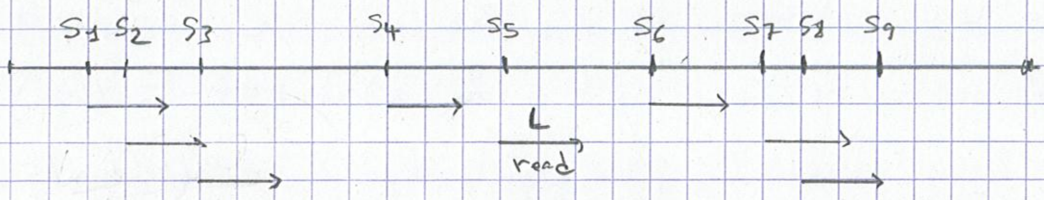
Article Roach

. Positionnement de lecture S_i sur un génome :

taille génome G



S_i : position de départ de la i^e lecture sur le génome

Les lectures sont générées aléatoirement et uniformément le long du génome (dont la longueur est G).

$D_i = S_{i+1} - S_i$: distance entre 2 positions de départ consécutives

- Dans l'article que je suis en train de lire, l'auteur prétend que la fonction de densité de probabilité (pdf) de D_i est une distribution beta (quand on normalise par la taille du génome, c'est-à-dire que les valeurs des S_i sont comprises [0-1]). Il indique :

$$\boxed{f_{D_k}(x) = n\left(1 - \frac{x}{G}\right)^{n-1}}$$

où n est le nombre de lectures générées

C'est effectivement une distribution beta avec $\alpha = 1$ $\beta = n$

Il a également supposé que les variables aléatoires sont continues

Il justifie la distri ci-dessus par : "The domain space of the spacings D_k is the surface of the simplex $D_0 + D_1 + D_2 + \ldots + D_n = G$ and their joint probability density is constant. This observation permits many probabilities of interest to be calculated by geometric considerations."

Comme je ne sais pas ce qu'il veut dire avec son histoire de simplex, je cherche à comprendre comment il arrive à ce résultat.

Je pars donc sur les statistiques d'ordre.

On s'intéresse donc aux statistiques d'ordre (2)

Soit X une va continue et $X_1, X_2, \dots X_n$ des v.a. iid selon la distribution de X : pdf $= f_X(x)$ et cdf $= F_X(x)$

On définit $X_{(1)}$ comme la plus petite des X_i précédents
$X_{(2)}$ comme la seconde des X_i — " "
etc

$X_{(1)}$ est le min des X_i, $X_{(n)}$ est le max des X_i

- On s'intéresse à l'événement $u < X_{(i)} < u+h$

Cet événement correspond à l'événement que $i-1$ va ont une valeur inf. à u, 1 va à une valeur dans $[u, u+h]$ et $n-i$ va ont une valeur $> u+h$

La proba de cet événement est une multinomiale où n essais sont distribués dans $k=3$ catégories avec les probabilités corresp.

$$P(u < X_{(i)} < u+h) = \frac{n!}{(i-1)!\,(n-i)!} F_X(u)^{i-1} f_X(u)\, h \left[1 - F_X(u+h)\right]^{n-i}$$

où on a supposé que $P(u < X_{(i)} < u+h) \simeq f_X(u)\, h$ quand h est petit.

Donc la densité de probabilité correspondante est :

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!\,(n-i)!} F_X(x)^{i-1} f_X(x) \left[1 - F_X(x)\right]^{n-i} \quad (1)$$

- Si X est distribuée de façon uniforme sur $(1-G)$

on a $f_X(x) = \frac{1}{G}$ $\quad F_X(x) = \frac{x}{G}$

Dans ce cas :

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!\,(n-i)!} \left(\frac{x}{G}\right)^{i-1} \frac{1}{G} \left[1 - \frac{x}{G}\right]^{n-i} \quad x \in [0-L]$$

- Si on pose $\frac{x}{G} = y$ $\quad \alpha = i \quad \beta = n-i+1 \qquad y \in [0-1)$

$$f_{X_{(i)}}(y) = \frac{1}{G} \frac{(\alpha+\beta-1)!}{(\alpha-1)!\,(\beta-1)!} y^{\alpha-1} (1-y)^{\beta-1} = \frac{1}{G} \text{beta}(i, n-i+1)$$

Si on considère maintenant la densité de probabilité jointe pour $S_{(i)}$ et $S_{(i+1)}$. On a de façon analogue à (1)

$$f_{X_{(i)}X_{(i+1)}}(x,y) = \frac{n!}{(i-1)!\,(n-i-1)!} \left(\frac{x}{G}\right)^{i-1} \frac{1}{G}\,\frac{1}{G} \left(1 - y/G\right)^{n-i-1} \qquad y > x$$

Evénement $D_i = d$ est le même que $S_{(i+1)} = x + d$ sachant que $S_{(i)} = x$. $\text{Proba}(D_i = d) = \text{Proba}(S_{(i+1)} = x+d \mid S_{(i)} = x)$

$$= \frac{\text{Proba}(S_{(i+1)} = x+d,\ S_{(i)} = x)}{\text{Proba}(S_{(i)} = x)}$$

donc

$$f_{D_i}(d) = \frac{n-i}{G} \frac{(1-x-d)^{n-i-1}}{(1-x)^{n-i}} = \frac{n-i}{G(1-x)} \left[\frac{1-x-d}{1-x}\right]^{n-i-1}$$

Il faut intégrer sur toutes les valeurs $x \in [0\ 1]$

- Arguments de Pyke (1965)

$X_1, X_2, \dots X_n$ iid random rv on $[0,1]$. Density function of $\underline{X} = (X_1, \dots, X_n)$ is

$$f_{\underline{X}}(x_1, \dots x_n) = \begin{cases} 1 & \text{if } 0 \le x_i \le 1 \text{ for } 1 \le i \le n \\ \emptyset & \text{otherwise} \end{cases}$$

$\underline{U} = (U_1, U_2, \dots, U_n)$ order statistics obtained by arranging the X_i's in increasing order

$$f_{\underline{U}}(u_1, \dots u_n) = \begin{cases} n! & \text{if } 0 \le u_1 \le u_2 \le \dots \le u_n \le 1 \\ \emptyset & \text{otherwise} \end{cases}$$

Set $U_\emptyset = \emptyset$ and $U_{n+1} = 1$. The spacing of the sample is $D_i = U_i - U_{i-1}$ $1 \le i \le n+1$

Since $d_1 + d_2 + \dots + d_{n+1} = 1$ the random vector $\underline{D} = (D_1, D_2, \dots D_{n+1})$ has a singular distribution but when restricted to this hyperplane has the density function

$$f_{\underline{D}}(d_1, d_2, \dots d_{n+1}) = \begin{cases} n! & \text{if } d_i \geqslant 0 \text{ and } d_1 + d_2 + \dots + d_{n+1} = 1 \\ \emptyset \end{cases} \quad (2)$$

Therefore most probability statements about $\underline{D}$ may be obtained by computing the volume of a subset of the plane $d_1 + d_2 + \dots + d_{n+1} = 1$ or simplex $\{d_i \geqslant 0$

From (2) the df of $\underline{D}$ remains unchanged under any permutation of its coordinates $\Rightarrow$ uniform spacings are interchangeable rv. $\Rightarrow$ the df of any spacing is equal to that of D_1 and the joint dF of any pair (D_i, D_j) is the same as (D_1, D_2)

$F_{D_i}(x) = F_{D_1}(x) = F_{U_1}(x) = 1 - (1-x)^n$ (integration of beta $(1, n)$ density

$$F_{D_i D_j}(x,y) = \Pr(U_1 \le x,\ U_2 - U_1 \le y) = n \int_0^x \left\{1 - \left(1 - \frac{y}{1-u}\right)^{n-1}\right\} (1-u)^{n-1}\, du$$