Infinitely-many-sites-Modell

Stochastische Modelle der Populationsbiologie, 13.1.2016

Definition 2.26 (Infinitely-many-sites-Modell (IMS), Watterson 1975)
Man nimmt an, dass jede Mutation eine neue, bisher noch nie mutierte Position am betrachteten Lokus betrifft.

Definition 2.26 (Infinitely-many-sites-Modell (IMS), Watterson 1975)
Man nimmt an, dass jede Mutation eine neue, bisher noch nie mutierte Position am betrachteten Lokus betrifft.

Mathematisch realisiert man dies z.B. folgendermaßen: Die betrachtete Stelle im Genom (eine gewisse Abfolge von Nukleotiden im DNS-Doppelstrang eines Chromosoms) entspricht $[0, 1]$, jede Mutation erhält eine neue, uniform aus $[0, 1]$ gewählte „Position“, der Typ eines Individuums ist ein (einfaches) Zählmaß auf $[0, 1]$ (bzw. alternativ eine Teilmenge von $[0, 1]$), der Typ eines Individuums gibt an, wo dieses relativ zu einen „Referenztyp“ (oder „Wildtyp“) mutiert ist.

Bemerkung. Das IMS-Modell (in der Literatur auch infinite-sites-Modell genannt) ist für viele praktische Zwecke eine angemessene Approximation für die Beschreibung von Mutationen auf dem Niveau der DNS-Sequenz : Wenn die Mutationsrate pro Basenpaar sehr klein und die betrachtete Stelle im Genom (der sog. Lokus) nicht „zu lang“ ist, ist es plausibel, die Möglichkeit der Mehrfachmutation einer Stelle (und andere Effekte, die im IMS-Modell nicht berücksichtigt werden, etwa Rekombination oder Insertionen/Deletionen längerer Stücke im Genom) zu vernachlässigen.

(Man denke an eine Abfolge von $L \gg 1$ Basenpaaren, bei Mutation wird eine der zufällig gewählte der L Positionen zufällig modifiziert.)

Beispiel 2.27 John Parsch, Colin D. Meiklejohn, and Daniel L. Hartl, Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of drosophila simulans, *Genetics* 159:647–657, (2001) berichten genetische Variabilität in einem ca. 1.700 Basenpaare langen Stück des Chromosoms 3 in einer (weltweiten) Stichprobe von 8 *Drosophila simulans* und einer Stichprobe von *Drosophila melanogaster*, zwei verwandten Arten von Taufliegen.

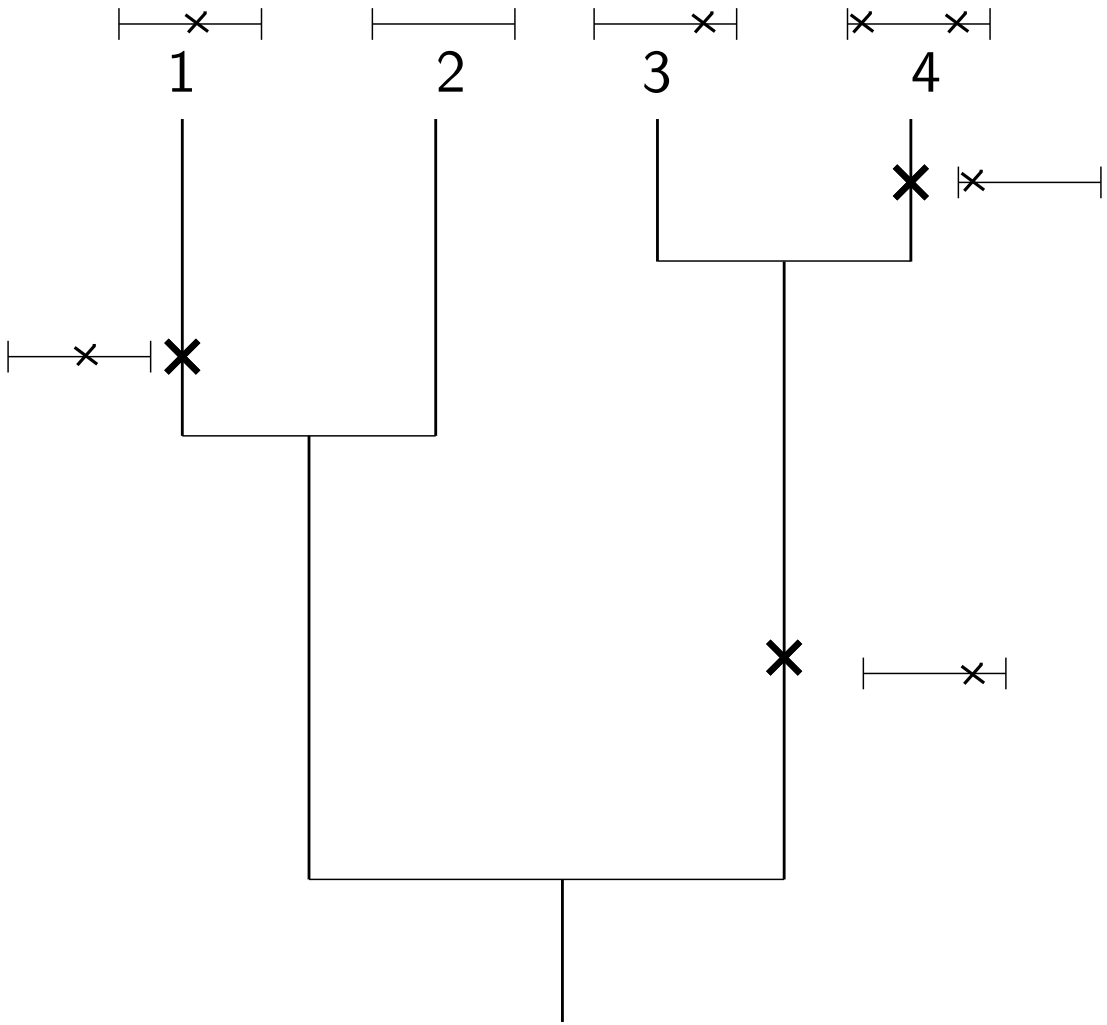
An insgesamt 31 Stellen sind Unterschiede zwischen den Individuen sichtbar.

	Position																															
	3	8	9	2	2	2	3	5	5	5	6	6	6	6	6	6	6	7	8	1	1	1	1	1	1	1	1	1	1	1	1	1
	5	3	3	4	9	9	6	3	4	8	4	5	5	5	5	6	6	0	4	3	3	9	1	3	0	2	3	8	2	4	6	7	8	9	6	8	1	2	6	4
s1	c	g	a	t	c	c	a	a	t	a	t	a	a	a	g	c	t	c	g	a	t	a	a	g	c	c	g	a	t	t	c	
s2	.	.	c	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.	
s3	a	c	.	c	a	t	g	c	c	c	g	g	g	g	a	t	c	t	a	t	c	c	t	c	t	g	t	t	g	c	a	
s4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
s5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.	
s6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
s7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.	
s8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.	
m1	.	.	.	c	a	t	g	c	c	c	a	g	.	.	.	.	.	t	.	.	c	c	.	.	t	g	.	t	g	c	a	

Tabelle: Beobachtete genetische Variabilität in einer Region in Chromosom 3 aus einer Stichprobe von 8 *Drosophila simulans* (Zeilen s1–s8) und einer Stichprobe von *Drosophila melanogaster* (Zeile m1) aus Parsch et al, *Genetics* 159:647–657, (2001). Siehe Figure 2 dort, wir betrachten hier nur den Teil der Sequenz, der die Gene *janA* und *janB* umfasst.

Aber: Parallelmutation an Position 644?

Modellvorstellung: n-Stichprobe entsteht aus n-Koaleszent, längs dessen Ästen sich mit Rate $\frac{\theta}{2}$ Mutationen ereignen (und jede trifft eine völlig neue Position)



Die Anzahl segregierender Stellen ist

$$S_n = \#\text{ verschiedene Mutationen, die in } n\text{-Stichprobe vorkommen}$$

(im Sinne von: Positionen, an denen sich mindestens zwei Stichproben unterscheiden).

Wenn $S_n = s$, so entsprechen die Beobachtungen einer $n \times s$-Datenmatrix $(D_{ik})_{i=1,\ldots,n;k=1,\ldots,s}$

$$D_{ik} = \mathbf{1}(\text{Stichprobe } i \text{ ist an } k\text{-ter segregierender Stelle mutiert}).$$

Die Beobachtungen aus Parsch et al als 0-1-Matrix

```
                                              1 1 1 1 1 1 1 1 1 1 1 1
          2 2 2 3 5 5 5 6 6 6 6 6 6 6 7     8 1 1 2 3 3 3 4 6 6 6 6 6
    3 8 9 4 9 9 6 3 4 8 4 5 5 5 5 6 6 0     4 3 3 9 0 2 8 5 1 7 8 8 9
    5 3 3 6 1 4 2 8 9 5 4 1 2 7 8 5 7 7     3 2 9 1 0 2 8 2 4 3 1 2 4
    -----------------------------------     -------------------------
s1  0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1     0 0 1 1 0 0 1 1 0 1 1 1 1
s2  0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1     0 0 1 1 0 0 1 0 0 1 1 1 1
s3  1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0     1 1 0 0 1 1 0 0 1 0 0 0 0
s4  0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1     0 0 1 1 0 0 1 1 0 1 1 1 1
s5  0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1     0 0 1 1 0 0 1 0 0 1 1 1 1
s6  0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1     0 0 1 1 0 0 1 1 0 1 1 1 1
s7  0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1     0 0 1 1 0 0 1 0 0 1 1 1 1
s8  0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1     0 0 1 1 0 0 1 0 0 1 1 1 1
```

Bemerkung (Unbekannter Wildtyp) Wenn man „nur“ die Stichprobe sieht und keine externen Zusatzinformationen (z.B. eine „outgroup“ durch inter-Spezies-Vergleich wie in obigem Bsp.) besitzt, kann man an den segregierenden Stellen nicht entscheiden, welcher Typ der Wildtyp und welcher die Mutante ist (im Genetik-Jargon: die Mutationen sind „unpolarisiert“).

Bemerkung (Unbekannter Wildtyp) Wenn man „nur“ die Stichprobe sieht und keine externen Zusatzinformationen (z.B. eine „outgroup“ durch inter-Spezies-Vergleich wie in obigem Bsp.) besitzt, kann man an den segregierenden Stellen nicht entscheiden, welcher Typ der Wildtyp und welcher die Mutante ist (im Genetik-Jargon: die Mutationen sind „unpolarisiert“).
In dieser Situation ist obige Datenmatrix nur bis auf „Umklappen“ von Spalten definiert, d.h. die eigentliche Information ist

$$S_n, \text{die Anzahl Mutationen („segregierende Stellen“)}$$

und

$$\Delta_{i,j}(k) = \begin{cases} 1, & \text{Stichproben } i \text{ und } j \text{ an } k\text{-ter segr. Stelle verschieden,} \\ 0, & \text{sonst} \end{cases}$$

für $k = 1, \ldots, S_n$.

Die Anzahl segregierender Stellen

$S_n = \#$ verschiedene Mutationen, die in n-Stichprobe vorkommen

Erinnerung: $\mathscr{L}_\theta(S_n) = \overset{n}{\underset{i=2}{*}} \mathrm{geom}(\frac{i-1}{\theta+i-1})$

(schreibe $S_n = S_{n,n} + S_{n,n-1} + \cdots + S_{n,2}$ mit $S_{n,j} = \#$ Mutationen, während Genealogie aus j Linien besteht)

Die Anzahl segregierender Stellen

$S_n = \#$ verschiedene Mutationen, die in n-Stichprobe vorkommen

Erinnerung: $\mathscr{L}_\theta(S_n) = \overset{n}{\underset{i=2}{*}} \operatorname{geom}(\frac{i-1}{\theta+i-1})$

(schreibe $S_n = S_{n,n} + S_{n,n-1} + \cdots + S_{n,2}$ mit $S_{n,j} = \#$ Mutationen, während Genealogie aus j Linien besteht)

Beobachtung 2.29.

$$\mathbb{E}_\theta[S_n] = \theta\, h_n, \quad \operatorname{Var}_\theta(S_n) = \theta\, h_n + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$$

mit $h_n = \sum_{i=1}^{n-1} \frac{1}{i}$.

$$\widehat{\theta}_W := \frac{S_n}{h_n}$$

ist ein erwartungtreuer Schätzer für θ ("Watterson-Schätzer"),

$\operatorname{Var}_\theta\left(\widehat{\theta}_W\right) \sim \frac{\theta}{h_n} \sim \frac{\theta}{\log n}$ für $n \to \infty$

(und $\widehat{\theta}_W$ ist asymptotisch normal)

Bemerkung 2.30 (Alternativer Zugang zu Beob. 2.29).
Wir könnten Erwartungswert und Varianz von S_n auch folgendermaßen berechnen: Gegeben die Gesamtlänge L_{ges} des Koaleszenten ist S_n $\text{Poi}\big((\theta/2)L_{\text{ges}}\big)$-verteilt.

$$L_{\text{ges}} \stackrel{d}{=} \sum_{j=2}^{n} j T_j,$$

mit $T_n, T_{n-1}, \ldots, T_2$ u.a., $\mathscr{L}(T_j) = \text{Exp}\big(\binom{j}{2}\big)$

Bemerkung 2.30 (Alternativer Zugang zu Beob. 2.29).
Wir könnten Erwartungswert und Varianz von S_n auch folgendermaßen berechnen: Gegeben die Gesamtlänge L_{ges} des Koaleszenten ist S_n $\text{Poi}\big((\theta/2)L_{\text{ges}}\big)$-verteilt.

$$L_{\text{ges}} \stackrel{d}{=} \sum_{j=2}^{n} j T_j,$$

mit $T_n, T_{n-1}, \ldots, T_2$ u.a., $\mathscr{L}(T_j) = \text{Exp}\big(\binom{j}{2}\big)$, somit

$$\begin{aligned}
\mathbb{E}_\theta[S_n] &= \frac{\theta}{2}\sum_{j=2}^{n} j\binom{j}{2} = \theta\sum_{i=1}^{n-1}\frac{1}{i}, \\
\text{Var}_\theta[S_n] &= \mathbb{E}_\theta\big[\text{Var}_\theta[S_n \mid L_{\text{ges}}]\big] + \text{Var}_\theta\big[\mathbb{E}_\theta[S_n \mid L_{\text{ges}}]\big] \\
&= \mathbb{E}_\theta\big[\frac{\theta}{2}L_{\text{ges}}\big] + \text{Var}_\theta\big[\frac{\theta}{2}L_{\text{ges}}\big] = \theta\sum_{i=1}^{n-1}\frac{1}{i} + \theta^2\sum_{i=1}^{n-1}\frac{1}{i^2}.
\end{aligned}$$

Satz 2.31 Eine untere Schranke für die Varianz von Schätzern für θ (Fu & Li, 1993):
Jeder erwartungstreue Schätzer für θ im IMS-Modell hat unter $\mathbb{P}_\theta$ Varianz $\geq \theta / \sum_{k=1}^{n-1} \frac{1}{\theta+k}$ ($\sim \theta / \log n$ für $n \to \infty$)

Satz 2.31 Eine untere Schranke für die Varianz von Schätzern für θ (Fu & Li, 1993):
Jeder erwartungstreue Schätzer für θ im IMS-Modell hat unter $\mathbb{P}_\theta$ Varianz $\geq \theta / \sum_{k=1}^{n-1} \frac{1}{\theta+k}$ ($\sim \theta / \log n$ für $n \to \infty$)
Nehmen wir an, wir könnten $S_{n,2} = s_{n,2}, \ldots, S_{n,n} = s_{n,2}$ beobachten (was anhand von Sequenzdaten an den Blättern des Koaleszenten nicht möglich ist)
Likelihoodfunktion

$$L_n(s_{n,2}, \ldots, s_{n,n}; \theta) = \prod_{j=2}^{n} \frac{j-1}{\theta+j-1} \left(\frac{\theta}{\theta+j-1} \right)^{s_{n,j}}$$

$$= (n-1)! \theta^{s_n} \prod_{j=2}^{n} (\theta+j-1)^{-(s_{n,j}+1)}$$

mit $s_n = s_{n,2} + \cdots + s_{n,n}$

$$\frac{\partial}{\partial \theta} \log L_n(s_{n,2}, \ldots, s_{n,n}; \theta) = \frac{s_n}{\theta} - \sum_{j=2}^{n} \frac{s_{n,j}+1}{\theta+j-1}$$

d.h. $\widehat{\theta}_{\mathrm{ML,hyp}}$ ist Lösung von $s_n = \theta \sum_{j=2}^{n} \frac{s_{n,j}+1}{\theta+j-1}$

(benutze Cramér-Rao-Ungleichung: $\mathrm{Var}_\theta(T(X)) \geq \left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)]\right)^2 / I(\theta)$)

$$\frac{\partial^2}{\partial\theta^2}\log L_n(s_{n,2},\ldots,s_{n,n};\theta) = -\frac{s_n}{\theta^2} + \sum_{j=2}^{n}\frac{s_{n,j}+1}{(\theta+j-1)^2}$$

Die Fisher-Information ist hier

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log L_n(s_{n,2},\ldots,s_{n,n};\theta)\right]$$

$$= \mathbb{E}_\theta\left[\frac{S_n}{\theta^2}\right] - \sum_{j=2}^{n}\mathbb{E}_\theta\left[\frac{S_{n,j}+1}{(\theta+j-1)^2}\right] = \frac{\sum_{k=1}^{n-1}1/k}{\theta} - \sum_{j=2}^{n}\frac{\theta+j-1}{(j-1)\,(\theta+j-1)^2}$$

$$= \frac{\sum_{k=1}^{n-1}1/k}{\theta} - \sum_{j=2}^{n}\frac{1}{(j-1)(\theta+j-1)} = \frac{1}{\theta}\sum_{k=1}^{n-1}\left(\frac{1}{k} - \frac{\theta}{k(\theta+k)}\right) = \frac{1}{\theta}\sum_{k=1}^{n-1}\frac{1}{\theta+k}$$

Definition 2.32: Frequenzspektrum (der segregrierenden Stellen)

$\xi_i^{(n)} = \#$ Mutationen, die in genau i der n Stichproben vorkommen
$(i = 1, \ldots, n-1)$
(Wir nehmen für den Moment an, dass an jeder Position der "Wildtyp" bekannt ist, z.B. durch Interspezies-Vergleich.)

Definition 2.32: Frequenzspektrum (der segregrierenden Stellen)

$\xi_i^{(n)} = \#$ Mutationen, die in genau i der n Stichproben vorkommen $(i = 1, \ldots, n-1)$

(Wir nehmen für den Moment an, dass an jeder Position der "Wildtyp" bekannt ist, z.B. durch Interspezies-Vergleich.)

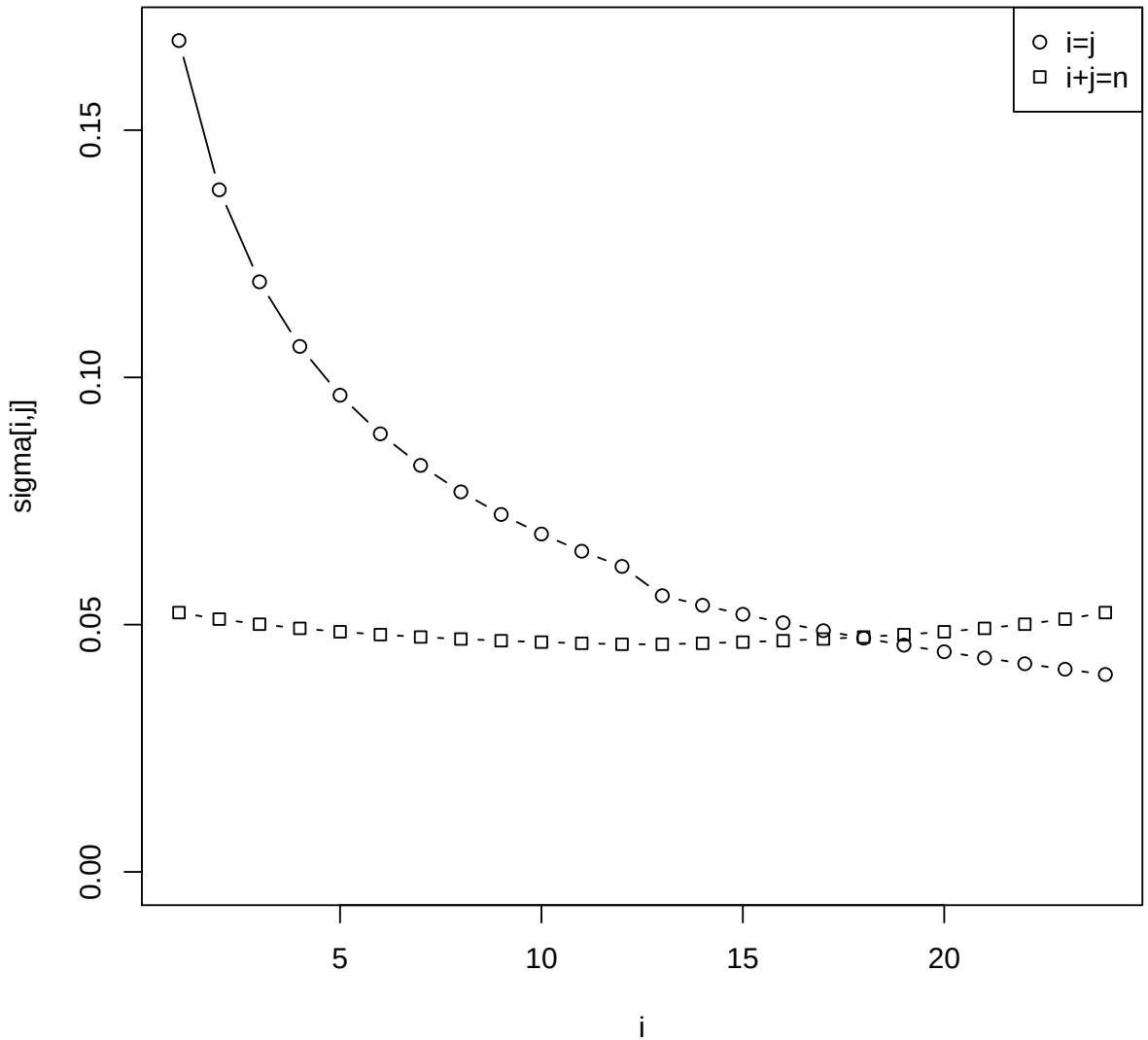
Satz 2.33 (Y.-X. Fu, 1995)

$$\mathbb{E}_\theta\left[\xi_i^{(n)}\right] = \frac{\theta}{i}, \quad \mathrm{Cov}_\theta\left(\xi_i^{(n)}, \xi_j^{(n)}\right) = \mathbf{1}_{i=j}\frac{\theta}{i} + \theta^2\sigma_{ij}, \quad 1 \le i \le j \le n$$

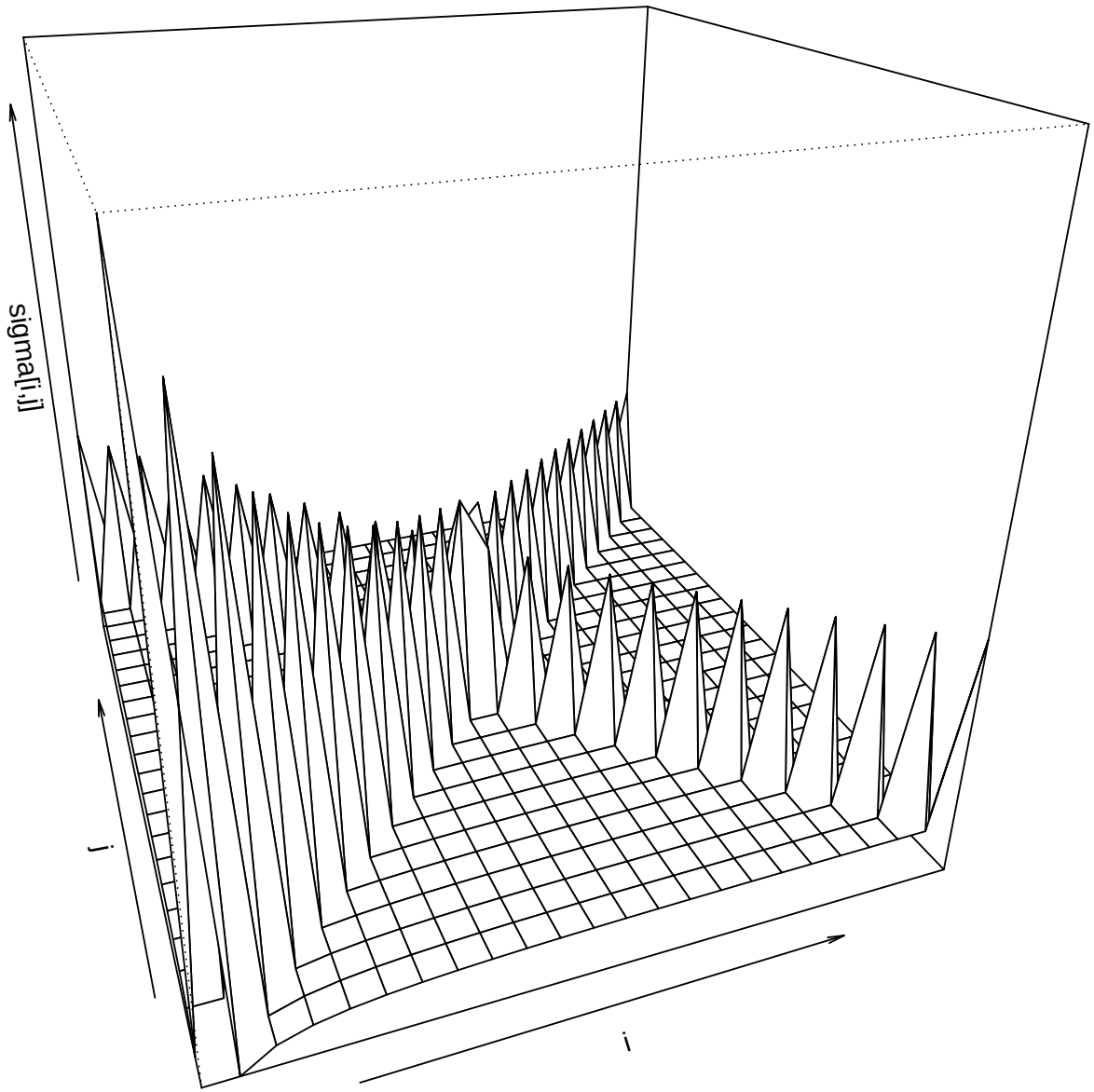
wobei mit $h_n := \sum_{i=1}^{n-1} \frac{1}{i}$, $\beta_n(i) := \frac{2n}{(n-i+1)(n-i)}\left(h_{n+1} - h_i\right) - \frac{2}{n-i}$

$$\sigma_{ii} = \begin{cases} \beta_n(i+1), & i < \frac{n}{2}, \\ 2\frac{h_n - h_i}{n-i} - \frac{1}{i^2}, & i = \frac{n}{2}, \\ \beta_n(i) - \frac{1}{i^2}, & i > \frac{n}{2}, \end{cases} \quad \sigma_{ij} = \begin{cases} \frac{\beta_n(i+1) - \beta_n(i)}{2}, & i + j < n, \\ \frac{h_n - h_i}{n-i} + \frac{h_n - h_j}{n-j} & \\ \quad - \frac{\beta_n(i) + \beta_n(j)}{2} - \frac{1}{ij}, & i + j = n, \\ \frac{\beta_n(j) - \beta_n(j+1)}{2} - \frac{1}{ij}, & i + j > n. \end{cases}$$

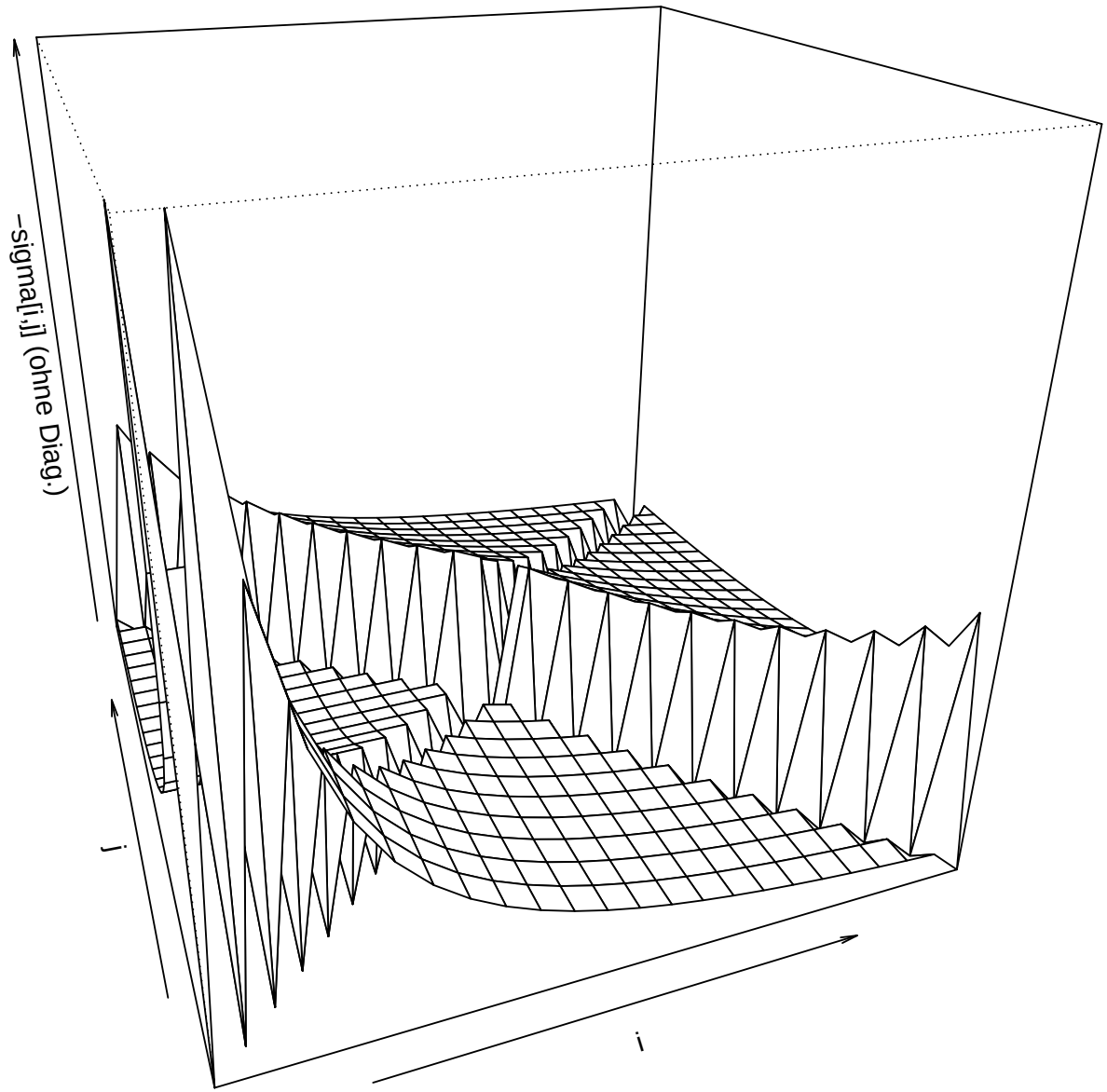
(s.a. Fu, loc. cit., Fig. 2, p. 188 und Fig. 3, p. 189)



$n = 25$, $\theta = 1$



$n = 25$, $\theta = 1$



$n = 25$, $\theta = 1$

Betrachte n-Stichprobe (im IMS-Modell)
Sei $\Delta_{ij} = \#$ Anzahl Mutationen, an denen sich Stichproben i und j unterscheiden $(1 \leq i < j \leq n)$

$\widehat{\theta}_\pi := \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \Delta_{ij}$ ("Tajimas $\widehat{\theta}_\pi$")

Betrachte n-Stichprobe (im IMS-Modell)
Sei $\Delta_{ij} = \#$ Anzahl Mutationen, an denen sich Stichproben i und j unterscheiden $(1 \leq i < j \leq n)$

$\widehat{\theta}_\pi := \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \Delta_{ij}$ ("Tajimas $\widehat{\theta}_\pi$")

Beob. 2.34 Es gebe s segregierende Stellen. $\widehat{\theta}_\pi = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \sum_{m=1}^{s} \mathbf{1}(\text{Stichpr. } i \text{ und } j \text{ unterschiedl. an } m\text{-ter segr. Stelle}) = \frac{1}{\binom{n}{2}} \sum_{k=1}^{n-1} \xi_k^{(n)} \, k(n-k)$ (mit $\xi_k^{(n)} = \#$ Mut., die in k Stichpr. vorkommen), d.h. $\widehat{\theta}_\pi$ ist eine Funktion des Frequenzspektrums.

Proposition 2.35 Es gilt $\mathbb{E}_\theta\left[\widehat{\theta}_\pi\right] = \theta$,

$\mathrm{Var}_\theta\left(\widehat{\theta}_\pi\right) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$.

(Insbesondere: $\widehat{\theta}_\pi$ ist erwartungstreuer Schätzer für θ, allerdings nicht konsistent: $\lim_{n\to\infty} \mathrm{Var}_\theta\left(\widehat{\theta}_\pi\right) = \frac{1}{3}\theta + \frac{2}{9}\theta^2 > 0$.)

Proposition 2.35 Es gilt $\mathbb{E}_\theta\left[\widehat{\theta}_\pi\right] = \theta$,
$\mathrm{Var}_\theta\left(\widehat{\theta}_\pi\right) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$.
(Insbesondere: $\widehat{\theta}_\pi$ ist erwartungstreuer Schätzer für θ, allerdings nicht konsistent: $\lim_{n\to\infty} \mathrm{Var}_\theta\left(\widehat{\theta}_\pi\right) = \frac{1}{3}\theta + \frac{2}{9}\theta^2 > 0$.)

Bem. $\frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} \Delta_{ij}$ wird in der Literatur auch mit π bezeichnet und die (empirische) "Nukleotid-Diversität" ("nucleotide diversity") genannt. ($\mathbb{E}_\theta[\pi] = \theta$ ist einer der Gründe für die Parametrisierung, dass Mutationen mit Rate $\theta/2$ längs der Genealogie erscheinen.)

Modell: Beobachtungen entsprechen Typen an den Blättern eines n-Koaleszenten im IMS-Modell

(biologische Interpretation: panmiktische, Population konstanter Größe, genetische Variabilität ist "neutral" (und es gibt keine Rekombination am betrachteten Lokus))

Frage: Passen beobachtete Sequenzdaten zum Modell?

Modell: Beobachtungen entsprechen Typen an den Blättern eines n-Koaleszenten im IMS-Modell

(biologische Interpretation: panmiktische, Population konstanter Größe, genetische Variabilität ist "neutral" (und es gibt keine Rekombination am betrachteten Lokus))

Frage: Passen beobachtete Sequenzdaten zum Modell?

Idee (F. Tajima, 1989): $\widehat{\theta}_W$ und $\widehat{\theta}_\pi$ sind beides erwartungstreue Schätzer für θ, d.h. wenn das Modell zutrifft, sollte

$$\widehat{\theta}_\pi - \widehat{\theta}_W \approx 0$$

bis auf "zufällige Fluktuationen".

Modell: Beobachtungen entsprechen Typen an den Blättern eines n-Koaleszenten im IMS-Modell

(biologische Interpretation: panmiktische, Population konstanter Größe, genetische Variabilität ist "neutral" (und es gibt keine Rekombination am betrachteten Lokus))

Frage: Passen beobachtete Sequenzdaten zum Modell?

Idee (F. Tajima, 1989): $\widehat{\theta}_W$ und $\widehat{\theta}_\pi$ sind beides erwartungstreue Schätzer für θ, d.h. wenn das Modell zutrifft, sollte

$$\widehat{\theta}_\pi - \widehat{\theta}_W \approx 0$$

bis auf "zufällige Fluktuationen".

Wie groß sind typische Fluktuationen, d.h. $\mathrm{Var}_\theta\left(\widehat{\theta}_\pi - \widehat{\theta}_W\right) = ?$

$$\mathrm{Var}_\theta\left(\widehat{\theta}_\pi - \widehat{\theta}_W\right) = ?$$

Bericht 2.36. Es gilt $\mathrm{Cov}_\theta\left(S_n, \widehat{\theta}_\pi\right) = \theta + \left(\frac{1}{2} + \frac{1}{n}\right)\theta^2$, also

$$\mathrm{Cov}_\theta\left(\widehat{\theta}_W, \widehat{\theta}_\pi\right) = \frac{\theta}{h_n} + \left(\frac{1}{2} + \frac{1}{n}\right)\frac{\theta^2}{h_n}$$

$\mathrm{Var}_\theta\left(\widehat{\theta}_\pi - \widehat{\theta}_W\right) = ?$

Bericht 2.36. Es gilt $\mathrm{Cov}_\theta\left(S_n, \widehat{\theta}_\pi\right) = \theta + \left(\frac{1}{2} + \frac{1}{n}\right)\theta^2$, also

$\mathrm{Cov}_\theta\left(\widehat{\theta}_W, \widehat{\theta}_\pi\right) = \frac{\theta}{h_n} + \left(\frac{1}{2} + \frac{1}{n}\right)\frac{\theta^2}{h_n}$

$\mathrm{Var}_\theta\left(\widehat{\theta}_\pi - \widehat{\theta}_W\right) = \left(\frac{n+1}{3(n-1)} - \frac{1}{h_n}\right)\theta + \left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{nh_n} + \frac{g_n}{h_n^2}\right)\theta^2$ und

$\widehat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1)$

(mit $\alpha_1 = \left(\frac{n+1}{3(n-1)} - \frac{1}{h_n}\right)/h_n$, $\alpha_2 = \left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{nh_n} + \frac{g_n}{h_n^2}\right)/(h_n^2 + g_n)$)

ist erwartungstreuer Schätzer für $\mathrm{Var}_\theta\left(\widehat{\theta}_\pi - \widehat{\theta}_W\right)$.

Tajimas D (F. Tajima, Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics (1989) 123: 585–595)

$$D := \frac{\widehat{\theta}_\pi - \widehat{\theta}_W}{\sqrt{\widehat{V}}}$$

($\widehat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1)$ mit $\alpha_1 = \left(\frac{n+1}{3(n-1)} - \frac{1}{h_n}\right)/h_n$,
$\alpha_2 = \left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{nh_n} + \frac{g_n}{h_n^2}\right)/(h_n^2 + g_n)$)
erfüllt $\mathbb{E}_\theta[D] \approx 0$, $\mathrm{Var}_\theta(D) \approx 1$.

Tajimas D (F. Tajima, Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics (1989) 123: 585–595)

$$D := \frac{\widehat{\theta}_\pi - \widehat{\theta}_W}{\sqrt{\widehat{V}}}$$

($\widehat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1)$ mit $\alpha_1 = \left(\frac{n+1}{3(n-1)} - \frac{1}{h_n}\right)/h_n$,
$\alpha_2 = \left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{nh_n} + \frac{g_n}{h_n^2}\right)/(h_n^2 + g_n)$)
erfüllt $\mathbb{E}_\theta[D] \approx 0$, $\mathrm{Var}_\theta(D) \approx 1$.

Kritische Werte?

$$D = \frac{\widehat{\theta}_\pi - \widehat{\theta}_W}{\sqrt{\widehat{V}}}$$

($\widehat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1)$ mit $\alpha_1 = \left(\frac{n+1}{3(n-1)} - \frac{1}{h_n}\right)/h_n$,
$\alpha_2 = \left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{nh_n} + \frac{g_n}{h_n^2}\right)/(h_n^2 + g_n)$)

Betrachte Ereignis $\{S_n = s\}$ (somit $\widehat{\theta}_W = s/h_n$, $\widehat{V} = \alpha_1 s + \alpha_2 s(s-1)$).
Kleinster möglicher Wert von $\widehat{\theta}_\pi$ ist $\frac{1}{\binom{n}{2}} s(n-1) = 2s/n$ (wenn $\xi_1^{(n)} + \xi_{n-1}^{(n)} = n$, $\xi_i^{(n)} = 0$ für $2 \le i \le n-2$), also kleinster mögl. Wert von D

$$\frac{2s/n - s/h_n}{\sqrt{\alpha_1 s + \alpha_2 s(s-1)}} \underset{s\to\infty}{\longrightarrow} \frac{2/n - 1/h_n}{\sqrt{\alpha_2}} =: d_{\min} \quad (= d_{\min}(n))$$

(und dies ist in einer "sternförmigen" Genealogie typisch)

$$D = \frac{\widehat{\theta}_\pi - \widehat{\theta}_W}{\sqrt{\widehat{V}}}$$

($\widehat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1)$ mit $\alpha_1 = \left(\frac{n+1}{3(n-1)} - \frac{1}{h_n}\right)/h_n$,
$\alpha_2 = \left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{nh_n} + \frac{g_n}{h_n^2}\right)/(h_n^2 + g_n)$)

Betrachte Ereignis $\{S_n = s\}$ (somit $\widehat{\theta}_W = s/h_n$, $\widehat{V} = \alpha_1 s + \alpha_2 s(s-1)$).
Größter möglicher Wert von $\widehat{\theta}_\pi$ ist
$\frac{1}{\binom{n}{2}} s\lceil n/2\rceil\lfloor n/2\rfloor = 2s\lceil n/2\rceil\lfloor n/2\rfloor/(n(n-1))$ (wenn $\xi^{(n)}_{\lceil n/2\rceil} = n$, $\xi^{(n)}_i = 0$ für $i \neq \lceil n/2\rceil$), also größter mögl. Wert von D

$$\frac{\frac{2s\lceil n/2\rceil\lfloor n/2\rfloor}{n(n-1)} - s/h_n}{\sqrt{\alpha_1 s + \alpha_2 s(s-1)}} \underset{s\to\infty}{\longrightarrow} \frac{\frac{2\lceil n/2\rceil\lfloor n/2\rfloor}{n(n-1)} - 1/h_n}{\sqrt{\alpha_2}} =: d_{\max} \ \ (= d_{\max}(n))$$

(und dies ist in einer "Hühnerbein"-Genealogie typisch)

Die exakte Verteilung von D unter der Nullhypothese

> "Beobachtungen entstehen durch die Typen an den Blättern eines n-Koaleszenten, längs dessen Kanten sich mit Rate $\theta/2$ Mutationen gemäß IMS-Modell ereignen"

ist nicht bekannt (und hängt von θ ab).

Die exakte Verteilung von D unter der Nullhypothese

> "Beobachtungen entstehen durch die Typen an den Blättern eines n-Koaleszenten, längs dessen Kanten sich mit Rate $\theta/2$ Mutationen gemäß IMS-Modell ereignen"

ist nicht bekannt (und hängt von θ ab).

Tajimas pragmatisch-heuristische Lösung: Approximiere die Vert. von D durch eine skalierte Beta-Verteilung, so dass der Träger $= [d_{\min}, d_{\max}]$, EW$= 0$ und Var$= 1$:

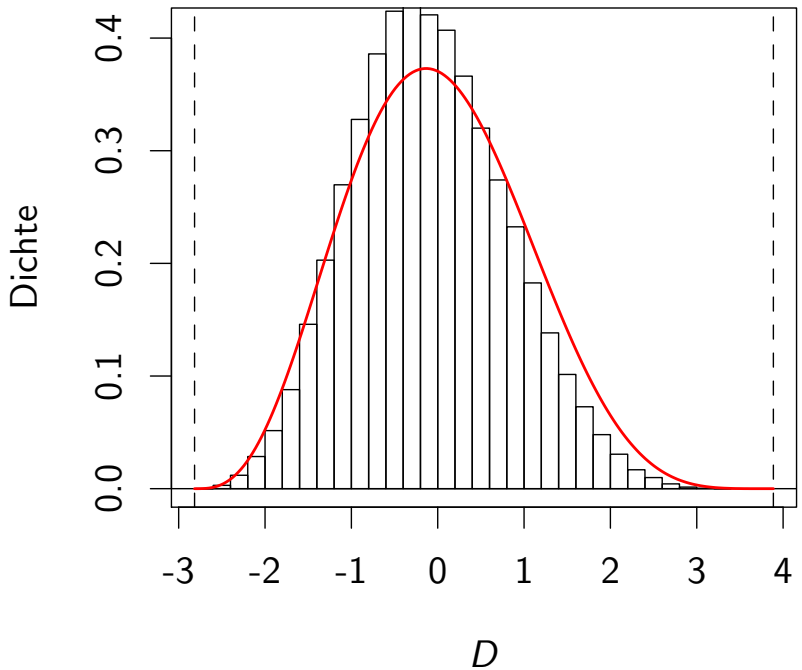
$$f_{\text{appr}}(d) = \frac{\Gamma(u+v)(d-d_{\min})^{u-1}(d_{\max}-d)^{v-1}}{\Gamma(u)\Gamma(v)(d_{\max}-d_{\min})^{u+v-1}}$$

mit

$$u = \frac{(1+d_{\max}d_{\min})d_{\min}}{d_{\max}-d_{\min}}, \quad v = -\frac{(1+d_{\max}d_{\min})d_{\max}}{d_{\max}-d_{\min}}.$$

d.h. wir betrachten den Ansatz $D \approx (d_{\max}-d_{\min})B + d_{\min}$ mit $B \sim \text{Beta}(u,v)$.

(Siehe z.B. F. Tajima, Genetics 123, Table 2, p. 592 für darauf fußende Konfidenzbereiche)



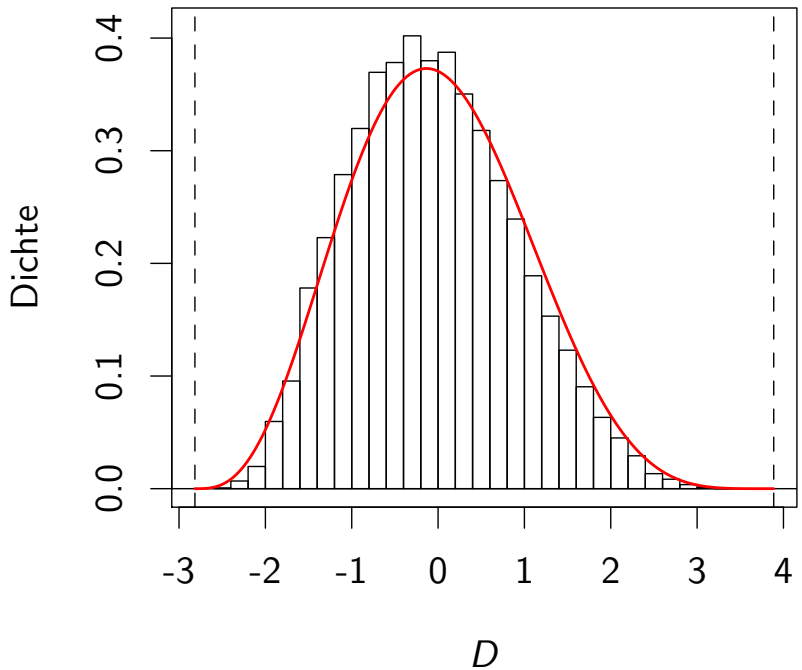


Abbildung: Simulation der Verteilung von D für $n = 25$ und $\theta = 10$ (links) bzw. $\theta = 2$ (rechts) unter dem Kingman-Koaleszenten mit IMS-Mutationen sowie angepasste skalierte Beta-Dichte. Histogramm jeweils basierend auf 100.000 simulierten Datensätzen.

Für die Daten aus Parsch et al (nur janA–janB) ergibt sich:
$n = 8$, $s = 31$, $\xi_1^{(8)} = 13$, $\xi_2^{(8)} = 1$, $\xi_7^{(8)} = 17$
$\widehat{\theta}_\pi = 7.93$,$\widehat{\theta}_W = 11.96$, $D = -1.79$

Tajimas Approximation liefert ein 95%-Konfidenzintervall für D unter dem Standard-Kingman-Koaleszenten von $[-1.663, 1.975]$ (s. Tajima, Genetics 123:585-595, (1989), Table 2),

d.h. die Abweichung von 0 ist auf dem 5%-Niveau signifikant.

Für die Daten aus Parsch et al (nur janA–janB) ergibt sich:
$n = 8$, $s = 31$, $\xi_1^{(8)} = 13$, $\xi_2^{(8)} = 1$, $\xi_7^{(8)} = 17$
$\widehat{\theta}_\pi = 7.93$,$\widehat{\theta}_W = 11.96$, $D = -1.79$

Tajimas Approximation liefert ein 95%-Konfidenzintervall für D unter dem Standard-Kingman-Koaleszenten von $[-1.663, 1.975]$ (s. Tajima, Genetics 123:585-595, (1989), Table 2),

d.h. die Abweichung von 0 ist auf dem 5%-Niveau signifikant.

Allerdings: Simonsen et als Ansatz (Genetics 141:413-429, (1995), Table 3 liefert (fuer $n = 10$, $S \in [27, 41]$) ein 95%-Konfidenzintervall für D unter dem Standard-Kingman-Koaleszenten von $[-1.80, 1.83]$.

Explizite Verteilung von S_n

Lemma 2.39 Es gilt

$$\mathbb{P}_\theta(S_n = m) = \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \left(\frac{\theta}{\theta+k}\right)^{m+1}, \quad m \in \mathbb{N}_0,$$

$$\mathbb{P}_\theta(S_n \leq s) = 1 - \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-1}{k} \left(\frac{\theta}{\theta+k}\right)^{s+1}, \quad s \in \mathbb{N}_0.$$

Explizite Verteilung von S_n

Lemma 2.39 Es gilt

$$\mathbb{P}_\theta(S_n = m) = \frac{n-1}{\theta}\sum_{k=1}^{n-1}(-1)^{k-1}\binom{n-2}{k-1}\left(\frac{\theta}{\theta+k}\right)^{m+1}, \quad m \in \mathbb{N}_0,$$

$$\mathbb{P}_\theta(S_n \le s) = 1 - \sum_{k=1}^{n-1}(-1)^{k-1}\binom{n-1}{k}\left(\frac{\theta}{\theta+k}\right)^{s+1}, \quad s \in \mathbb{N}_0.$$

$S_n = S_{n,n} + S_{n,n-1} + \cdots + S_{n,2}$ mit $S_{n,j} \sim \mathrm{geom}(\frac{i-1}{\theta+i-1})$ u.a.

Sei $u \in [0,1]$: $\mathbb{E}\left[u^{S_{n,j}}\right] = \sum_{\ell=0}^{\infty} u^\ell \frac{j-1}{\theta+j-1}\left(\frac{\theta}{\theta+j-1}\right)^\ell = \frac{j-1}{\theta+j-1}\frac{1}{1-u\frac{\theta}{\theta+j-1}} = \frac{j-1}{j-1+\theta(1-u)}$,

$\mathbb{E}\left[u^{S_n}\right] = \prod_{j=2}^{n}\mathbb{E}\left[u^{S_{n,j}}\right] = \prod_{k=1}^{n-1}\frac{k}{k+\theta(1-u)}$

und $\prod_{k=1}^{n-1}\frac{k}{k+z} = \sum_{k=1}^{n-1}\frac{a_{n,k}}{k+z}$ $(z \in \mathbb{C}\setminus -\mathbb{N})$ mit

$a_{n,k} = \frac{(n-1)!}{\prod_{j\neq k}^{n-1}(j-k)} = (-1)^k(n-1)\binom{n-2}{k-1}$

also $\mathbb{E}_\theta\left[u^{S_n}\right] = \sum_{m=0}^{\infty}u^m\mathbb{P}_\theta(S_n=m) = \sum_{k=1}^{n-1}a_{n,k}\sum_{m=0}^{\infty}\left(\frac{\theta}{\theta+k}\right)^m u^m =$

$\sum_{m=0}^{\infty}u^m\sum_{k=1}^{n-1}a_{n,k}\left(\frac{\theta}{\theta+k}\right)^m$

$$\mathbb{P}_\theta\left(S_n \leq s\right) = \sum_{m=0}^{s} \mathbb{P}_\theta\left(S_n = m\right) = \sum_{m=0}^{s} \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \left(\frac{\theta}{\theta+k}\right)^{m+1}$$

$$= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \sum_{m=0}^{s} \left(\frac{\theta}{\theta+k}\right)^{m+1}$$

$$= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \frac{\theta}{\theta+k} \frac{1-\left(\frac{\theta}{\theta+k}\right)^{s+1}}{1-\left(\frac{\theta}{\theta+k}\right)}$$

$$= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \frac{\theta}{k} \left(1-\left(\frac{\theta}{\theta+k}\right)^{s+1}\right)$$

$$= \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-1}{k} \left(1-\left(\frac{\theta}{\theta+k}\right)^{s+1}\right)$$

$$= 1 - \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-1}{k} \left(\frac{\theta}{\theta+k}\right)^{s+1}$$

(denn $-\sum_{k=1}^{n-1}(-1)^k \binom{n-1}{k} = 1 - (1-1)^{n-1} = 1$)