

Infinitely-many-sites-Modell: Jede Mutation trifft eine neue Position am betrachteten Locus

Modellvorstellung:  $n$ -Stichprobe entsteht aus  $n$ -Koaleszent, längs dessen Ästen sich mit Rate  $\frac{\theta}{2}$  Mutationen ereignen

[Bild an der Tafel]

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 Bp langen Stück des Chromosoms 3 in einer (weltweiten) Stichprobe von 8 *Drosophila simulans* und einer Stichprobe von *Drosophila melanogaster*

			1 1 1 1 1 1 1 1 1 1 1 1	1 1 2 2 2 2 2 2 2 2 2 2 2 2
	2 2 2 3 5 5 5 6 6 6 6 6 6 7	8 1 1 2 3 3 3 4 6 6 6 6 6	8 9 0 0 0 0 0 1 1 1 1 1 1 1	
	3 8 9 4 9 9 6 3 4 8 4 5 5 5 6 6 0	4 3 3 9 0 2 8 5 1 7 8 8 9	3 6 0 0 1 2 2 5 6 7 7 7 8 9 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	6 8 1 6 4 0 8 7 4 1 6 9 8 4 5	
	-----	-----	-----	
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	a c g t c t g a t a a g c g -	
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 . . . . . . . .	g . c . t g a g g g c c - t t	
s6	. . . . . . . . . . . . . . . . . . . .	. . . . . . . . . . . . . . . . . . . .	. . . . . . . . . . . . . . . . . .	
s7	. . . . . . . . . . . . . . . . . . . .	. . . . . . . . . g . . . . . . . .	g . c . t g a g g g c c - t t	
s8	. . . . . . . . . . . . . . . . . . . .	. . . . . . . . . g . . . . . . . .	. t . c . g . g g g c c . . . .	
m1	. . . c a t g c c c a g . . . . . t	. . . c c . . t g . t g c a	g . . . . g . g g g c c - t t	
	-----	-----	-----	
	janA	janB	ocn	

Insgesamt: 44 SNPs und 2 indels

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

		1 1 1 1 1 1 1 1 1 1 1 1	1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
	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 6	8 9 0 0 0 0 0 0 1 1 1 1 1 1 1 1
	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 8 9	3 6 0 0 1 2 2 5 6 7 7 7 8 9 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	6 8 1 6 4 0 8 7 4 1 6 9 8 4 5
	-----	-----	-----
s1	0 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 1 0 1 1 1 1 1	1 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1
s2	0 0 1 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 1	1 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1
s3	1 1 0 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 0	1 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1
s4	0 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 1 0 1 1 1 1 1	1 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1
s5	0 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 1	0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0
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 1	1 0 0 0 0 1 0 1 1 1 1 1 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 1	0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0
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 1	1 1 0 1 0 0 0 0 0 0 0 0 0 1 1 1
	-----	-----	-----
	janA	janB	ocn

Beachte: Es gibt Inkonsistenzen zum IMS-Modell: Mehrfachmutation an Stelle 644, inkompatible Mutationsmuster zwischen janB und ocn (Rekombination?)

[Wir werden uns später auf den Bereich janA–janB einschränken.]

## Die Anzahl segregierender Stellen

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

Erinnerung:  $\mathcal{L}_\theta(S_n) = \prod_{i=2}^n \text{geom}\left(\frac{i-1}{\theta+i-1}\right)$

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

[Skizze an Tafel]

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

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

$\hat{\theta}_W := \frac{S_n}{h_n}$  ist ein erwartungstreuer Schätzer für  $\theta$  (“Watterson-Schätzer”),

$$\text{Var}_\theta\left(\hat{\theta}_W\right) \sim \frac{\theta}{h_n} \sim \frac{\theta}{\log n} \text{ für } n \rightarrow \infty$$

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

# Explizite Verteilung von $S_n$

**Beobachtung 16** 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 $S_n$

**Beobachtung 16** Es gilt

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

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

$S_n = S_{n,n} + S_{n,n-1} + \dots + S_{n,2}$  mit  $S_{n,j} \sim \text{geom}\left(\frac{j-1}{\theta+j-1}\right)$  u.a.

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

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

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

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

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

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

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

Eine untere Schranke für die Varianz von Schätzern für  $\theta$  (Fu & Li, 1993):  
Jeder erwartungstreue Schätzer für  $\theta$  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  $n \rightarrow \infty$ )



Eine untere Schranke für die Varianz von Schätzern für  $\theta$  (Fu & Li, 1993):  
 Jeder erwartungstreue Schätzer für  $\theta$  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  $n \rightarrow \infty$ )

Nehmen wir an, wir könnten  $S_{n,2} = s_{n,2}, \dots, S_{n,n} = s_{n,2}$  beobachten (was anhand von Sequenzdaten an den Blättern des Koaleszenten nicht möglich ist)  
 Likelihoodfunktion

$$\begin{aligned} L_n(s_{n,2}, \dots, s_{n,n}; \theta) &= \prod_{j=2}^n \frac{j-1}{\theta+j-1} \left( \frac{\theta}{\theta+j-1} \right)^{s_{n,j}} \\ &= (n-1)! \theta^{s_n} \prod_{j=2}^n (\theta+j-1)^{-(s_{n,j}+1)} \end{aligned}$$

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

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

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

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

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

Die Fisher-Information ist hier

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta \left[ \frac{\partial^2}{\partial \theta^2} \log L_n(s_{n,2}, \dots, s_{n,n}; \theta) \right] \\ &= \mathbb{E}_\theta \left[ \frac{S_n}{\theta^2} \right] - \sum_{j=2}^n \mathbb{E}_\theta \left[ \frac{s_{n,j} + 1}{(\theta + j - 1)^2} \right] = \frac{\sum_{k=1}^{n-1} 1/k}{\theta} - \sum_{j=2}^n \frac{\theta + j - 1}{(j - 1)(\theta + j - 1)^2} \\ &= \frac{\sum_{k=1}^{n-1} 1/k}{\theta} - \sum_{j=2}^n \frac{1}{(j - 1)(\theta + j - 1)} = \frac{1}{\theta} \sum_{k=1}^{n-1} \left( \frac{1}{k} - \frac{\theta}{k(\theta + k)} \right) = \frac{1}{\theta} \sum_{k=1}^{n-1} \frac{1}{\theta + k} \end{aligned}$$

**Erinnerung:** Frequenzspektrum (der segregierenden Stellen)

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

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

**Satz 15** (Y.-X. Fu, *Theor. Pop. Biol.* 48, 172–197 (1995))

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

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

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

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

Betrachte  $n$ -Stichprobe (im IMS-Modell)

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

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

**Beob.** Es gebe  $s$  segregierende Stellen.  $\widehat{\theta}_\pi =$

$$\frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \sum_{m=1}^s \mathbf{1}(\text{Stichpr. } i \text{ und } j \text{ unterschiedl. an } m\text{-ter segr. Stelle}) =$$

$$\frac{1}{\binom{n}{2}} \sum_{k=1}^{n-1} \xi_k^{(n)} k(n-k) \text{ (mit } \xi_k^{(n)} = \# \text{ Mut., die in } k \text{ Stichpr. vorkommen),}$$

d.h.  $\widehat{\theta}_\pi$  ist eine Funktion des Frequenzspektrums.

**Proposition 17.** Es gilt  $\mathbb{E}_\theta [\widehat{\theta}_\pi] = \theta$ ,  $\text{Var}_\theta (\widehat{\theta}_\pi) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$ .

(Insbesondere:  $\widehat{\theta}_\pi$  ist erwartungstreuer Schätzer für  $\theta$ , allerdings nicht

konsistent:  $\lim_{n \rightarrow \infty} \text{Var}_\theta (\widehat{\theta}_\pi) = \frac{1}{3}\theta + \frac{2}{9}\theta^2 > 0$ .)

**Bem.**  $\frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \Delta_{ij}$  wird in der Literatur auch mit  $\pi$  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 Locus))

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 Locus))

Frage: Passen beobachtete Sequenzdaten zum Modell?

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

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

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 Locus))

Frage: Passen beobachtete Sequenzdaten zum Modell?

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

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

bis auf “zufällige Fluktuationen”.

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

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

$$\text{(Erinnerung: } \text{Var}_\theta \left( \hat{\theta}_\pi \right) = \frac{n+1}{3(n-1)} \theta + \frac{2(n^2+n+3)}{9n(n-1)} \theta^2, \text{Var}_\theta \left( \hat{\theta}_W \right) = \frac{\theta}{h_n} + \frac{\theta^2 g_n}{h_n^2}$$

$$\text{mit } h_n = \sum_{i=1}^{n-1} \frac{1}{i}, \quad g_n = \sum_{i=1}^{n-1} \frac{1}{i^2}$$

$$\text{Es gilt } \text{Cov}_\theta \left( S_n, \hat{\theta}_\pi \right) = \theta + \left( \frac{1}{2} + \frac{1}{n} \right) \theta^2, \text{ also}$$

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

**Lemma.**

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

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

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

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



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

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

$(\hat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1))$  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  $\mathbb{E}_\theta[D] \approx 0$ ,  $\text{Var}_\theta(D) \approx 1$ .

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

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

( $\hat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1)$  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  $\mathbb{E}_\theta[D] \approx 0$ ,  $\text{Var}_\theta(D) \approx 1$ .

Kritische Werte?

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

( $\hat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1)$  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  $\hat{\theta}_W = s/h_n$ ,  $\hat{V} = \alpha_1 s + \alpha_2 s(s-1)$ ).

Kleinster möglicher Wert von  $\hat{\theta}_\pi$  ist  $\frac{1}{\binom{n}{2}} s(n-1) = 2s/n$  (wenn  $\xi_1^{(n)} = n$ ,

$\xi_i^{(n)} = 0$  für  $2 \leq i \leq n$ ), also kleinster mögl. Wert von  $D$

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

(und dies ist in einer “sternförmigen” Genealogie typisch)

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

( $\hat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1)$ ) 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  $\hat{\theta}_W = s/h_n$ ,  $\hat{V} = \alpha_1 s + \alpha_2 s(s-1)$ ).

Größter möglicher Wert von  $\hat{\theta}_\pi$  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  $\xi_{\lceil n/2 \rceil}^{(n)} = n$ ,  $\xi_i^{(n)} = 0$  für  $i \neq \lceil n/2 \rceil$ ), also größter mögl. Wert von  $D$

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

(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  $\theta$  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  $\theta$  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 :

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

mit

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

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

Bem.: Simonson et al, Genetics 141:413–429 (1995) geben eine “exakte” Lösung.

Für die Daten aus Parsch et al (nur janA–janB) ergibt sich:

$$n = 8, s = 31, \xi_1^{(8)} = 13, \xi_2^{(8)} = 1, \xi_7^{(8)} = 17$$

$$\hat{\theta}_\pi = 7.93, \hat{\theta}_W = 11.96, D = -1.79$$

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:

$$n = 8, s = 31, \xi_1^{(8)} = 13, \xi_2^{(8)} = 1, \xi_7^{(8)} = 17$$

$$\hat{\theta}_\pi = 7.93, \hat{\theta}_W = 11.96, D = -1.79$$

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  $[-1.80, 1.83]$ .