

Matematiska och statistiska metoder för genletning

Ola Hössjer

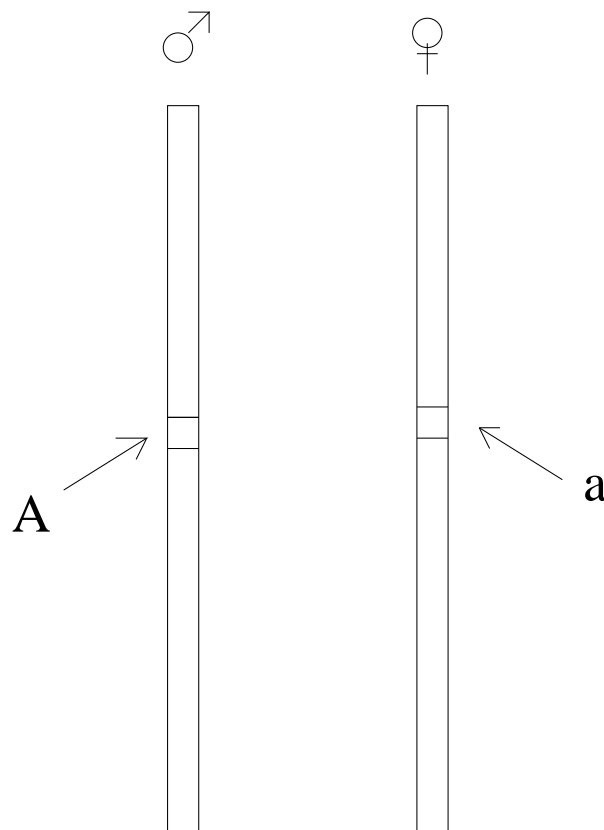
Matematiska institutionen
Avd. för matematisk statistik
Stockholms universitet

Mänskligt DNA

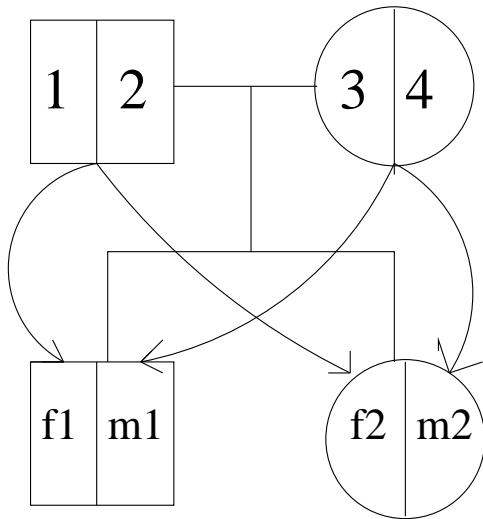
- 22 par av autosomer
- 2 könskromosomer
- 23 kromosomer nedärvda från fader, 23 från moder
- $3 \cdot 10^9$ baspar
- $1 \cdot 10^6 - 10 \cdot 10^6$ baspar polymorfa (olika mellan individer)

Gener och alleler

- Gen: Sekvens av ca 1000-10000 baspar - kodar för protein
- Olika former av en gen orsakade av t ex mutationer i enstaka baspar.
- Vissa former av genen sjukdomsalstrande (A), andra normala (a).



Kärnfamilj



1=DNA från farfar
 2=DNA från farmor
 3=DNA från morfar
 4=DNA från mormor

f1 och f2 från 1 eller 2
 m1 och m2 från 3 eller 4

□ =man
 ○ =kvinna

Z = antal DNA-bitar syskonen ärver från samma mor-/farförälder vid given kromosomposition

$f1 = f2?$	$m1 = m2?$	Z
Nej	Nej	0
Nej	Ja	1
Ja	Nej	1
Ja	Ja	2

Med $f1 = f2$ menas att $f1$ och $f2$ ärvts ner från samma farförälder, och motsv för $m1 = m2$.

DNA delat av ett syskonpar

Antagande: DNA registreras överallt hos (alla) familjemedlemmar

$Z(x)$ = antal DNA-bitar som syskonen ärvt ner från samma mor/farförälder i position x längs en kromosom

Mendels ärftlighetslagar ger¹

$Z(x) \in \text{Bin}(2, 0.5)$
= antal krona vid två slantsinglingar med symmetriskt mynt,

↓

$$P(Z(x) = 0) = 0.25$$

$$P(Z(x) = 1) = 0.5$$

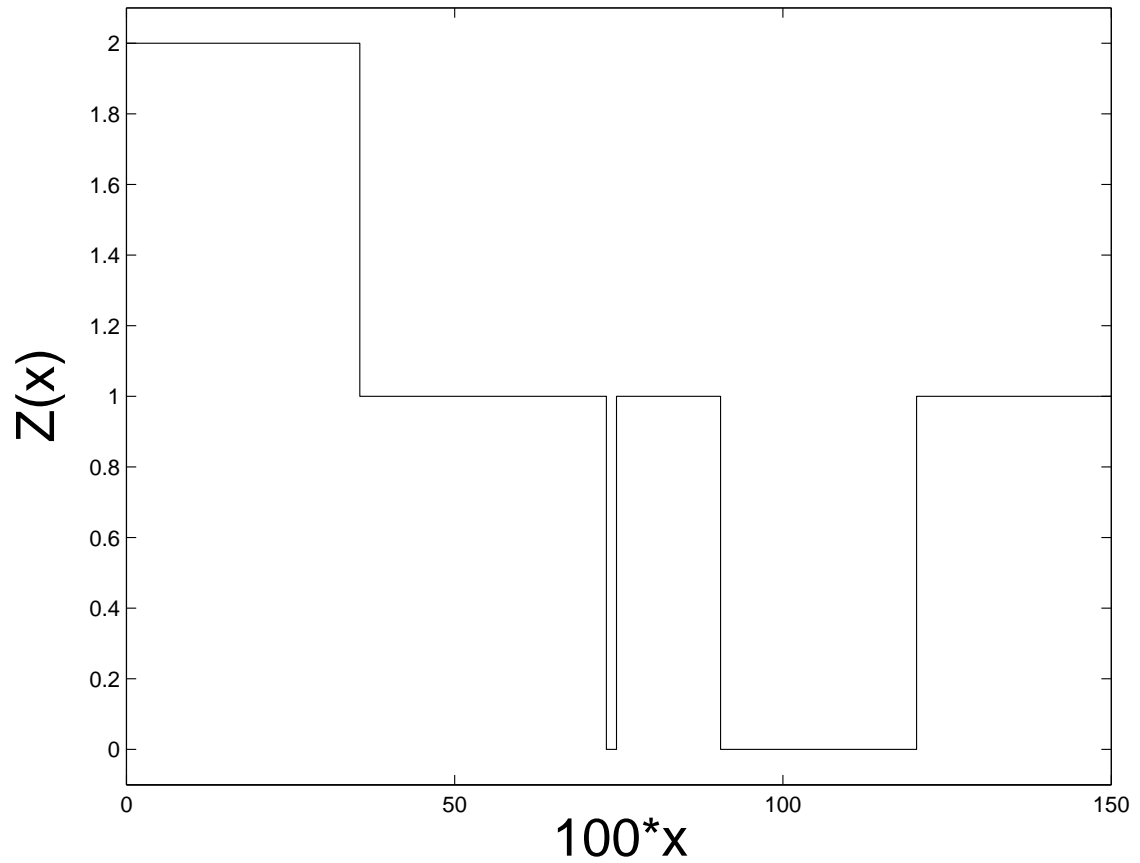
$$P(Z(x) = 2) = 0.25$$

↓

I genomsnitt hälften av syskonens DNA från samma mor-/farförälder.

¹slantsingling = nedärvning, krona = nedärvning från samma mor-/farförälder

Plot av $Z(x)$, ett syskonpar

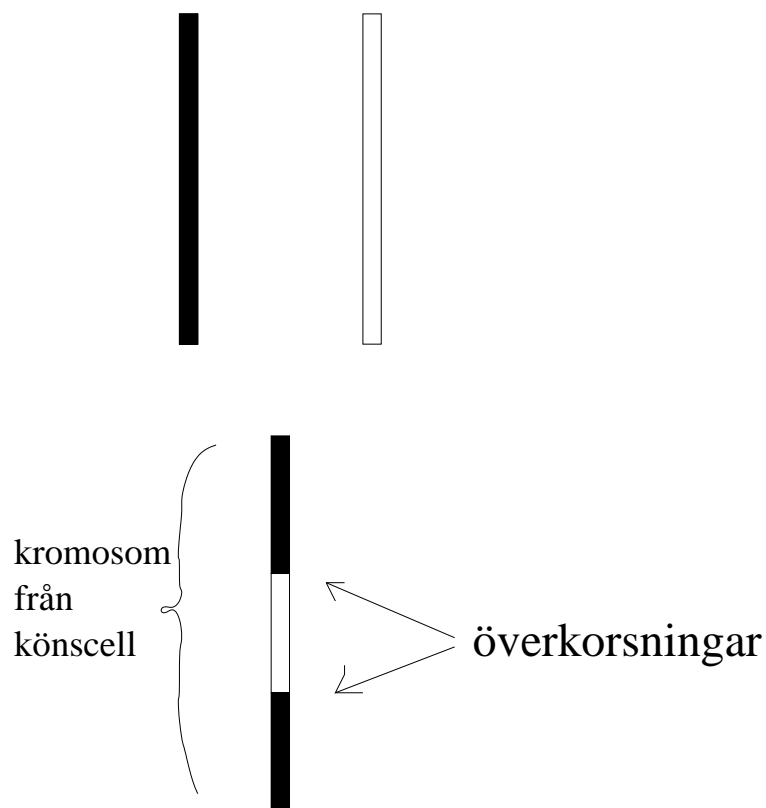


Mängden delat DNA varierar på grund av överkorsningar.

Överkorsningar

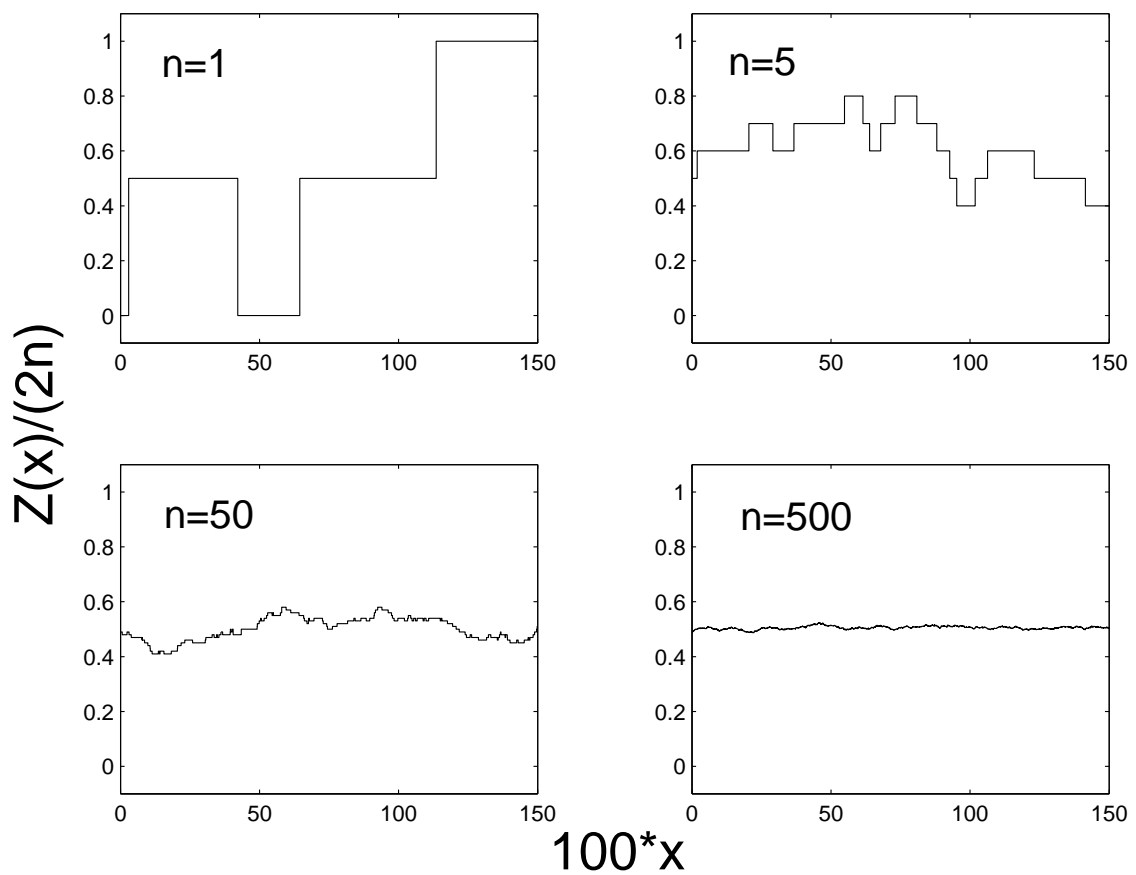
Slumpmässiga punkter längs kromosomen där nedärvning växlar från mormor till morfar (eller från farmor till farfar).

- Intensitet genomsnitt 1 överkorsning/ 10^8 baspar
- 1.5 överkorsningar per kromosom av genomsnittslängd
- Intensitet varierar med kromosomposition
- Fler överkorsningar vid bildande av ägg- än vid spermceller.



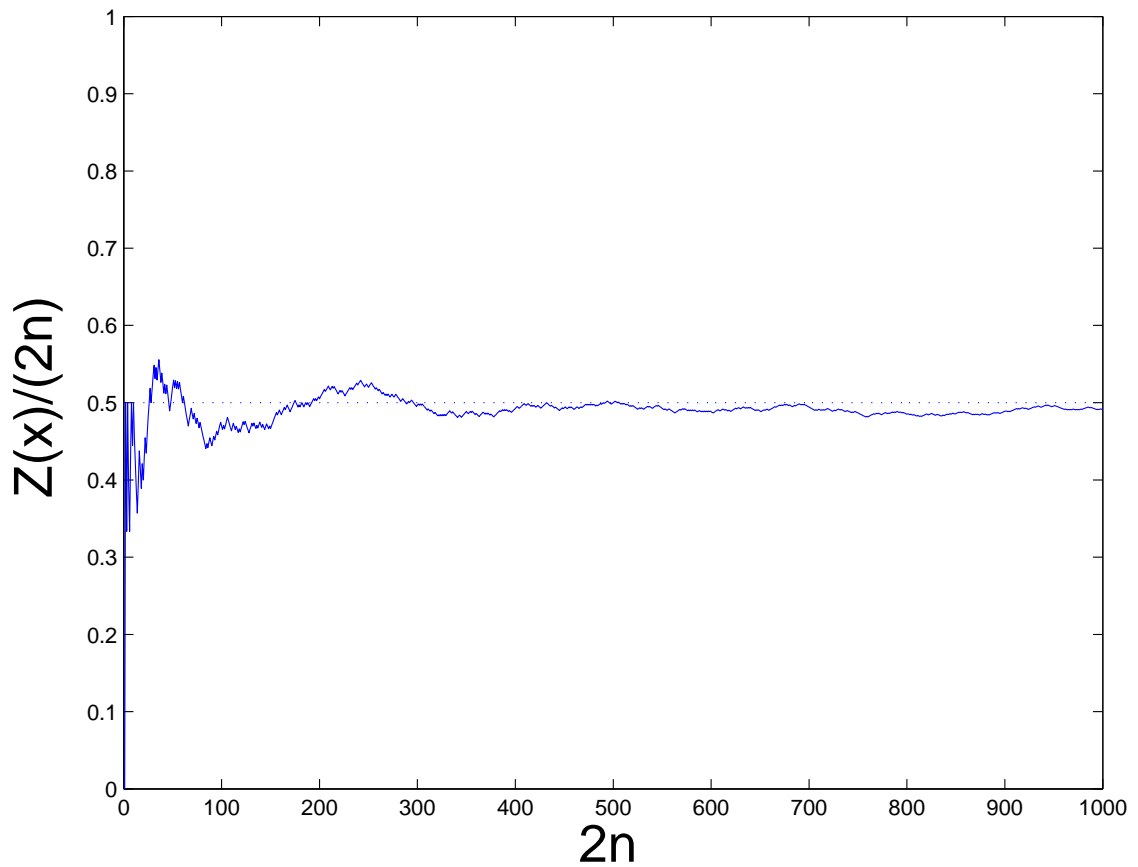
n syskonpar

- $Z(x) \in \text{Bin}(2n, 0.5)$
= totalt antal DNA-bitar från samma mor-/
farföräldrar i position x för n syskonpar
= antal krona vid $2n$ slantsinglingar med
symmetriskt mynt



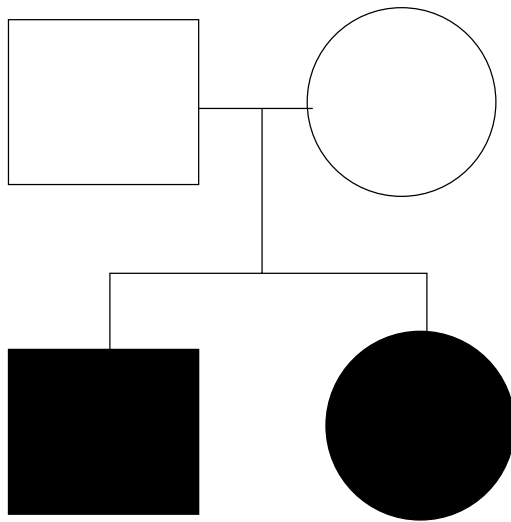
Stora talens lag

$$\begin{aligned} Z(x) &\in \text{Bin}(2n, 0.5) \\ &\Downarrow \\ Z/(2n) &\rightarrow 0.5 \text{ d\aa } n \rightarrow \infty. \end{aligned}$$



$$P(Z(x) = k) = \binom{2n}{k} 2^{-2n}, \quad k = 0, 1, \dots, 2n$$

Affekterat syskonpar



fylld symbol =
affekterad

Anta

- a = normal form av gen
- A = sjukdomsalstrande form av gen
- q = $P(A)$
- f = sjukdomsrisk, normalperson
- ψ = relativ risk för A jämfört med a

Dvs

$$\begin{aligned}P(\text{sjuk}|aa) &= f, \\P(\text{sjuk}|Aa) &= f\psi, \\P(\text{sjuk}|AA) &= f\psi^2.\end{aligned}$$

Affekterat syskonpar, forts

Man kan visa:

$$Z(\tau) \in \text{Bin}(2n, p),$$

där

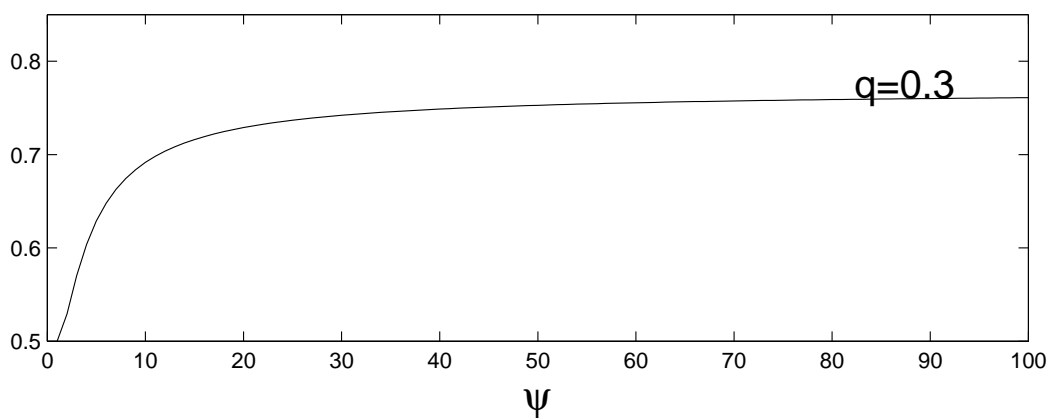
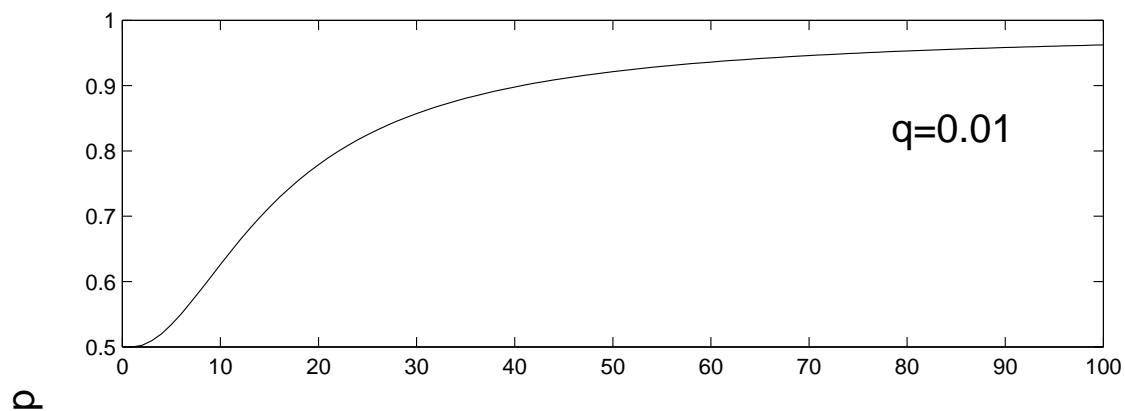
$$\tau = \text{position för sjukdomsgen}$$

$$\begin{aligned} p &= \frac{1+q(\psi^2-1)}{1+q(\psi^2-1)+(1+q(\psi-1))^2} \\ &= \text{proportionen DNA från samma} \\ &\quad \text{mor-/farförälder vid } \tau \end{aligned}$$

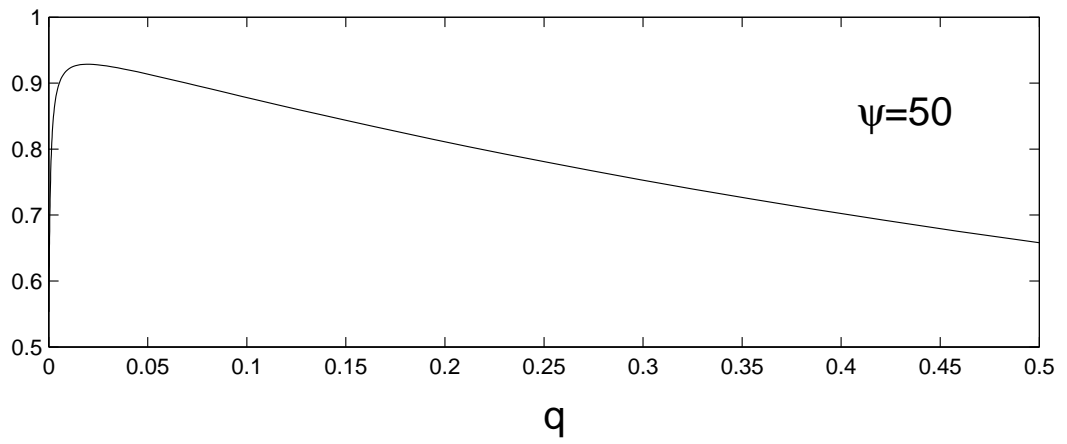
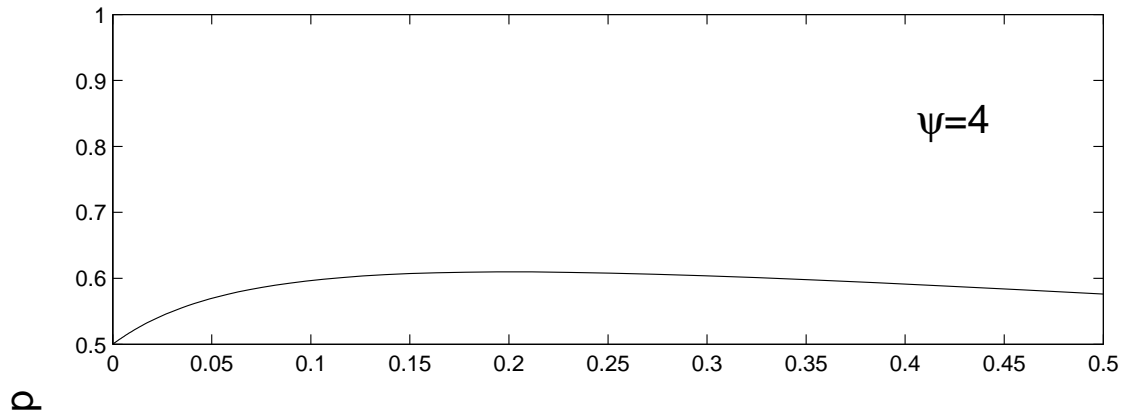
Dvs

$$\psi > 1 \implies p > 0.5.$$

p som funktion av ψ



p som funktion av q



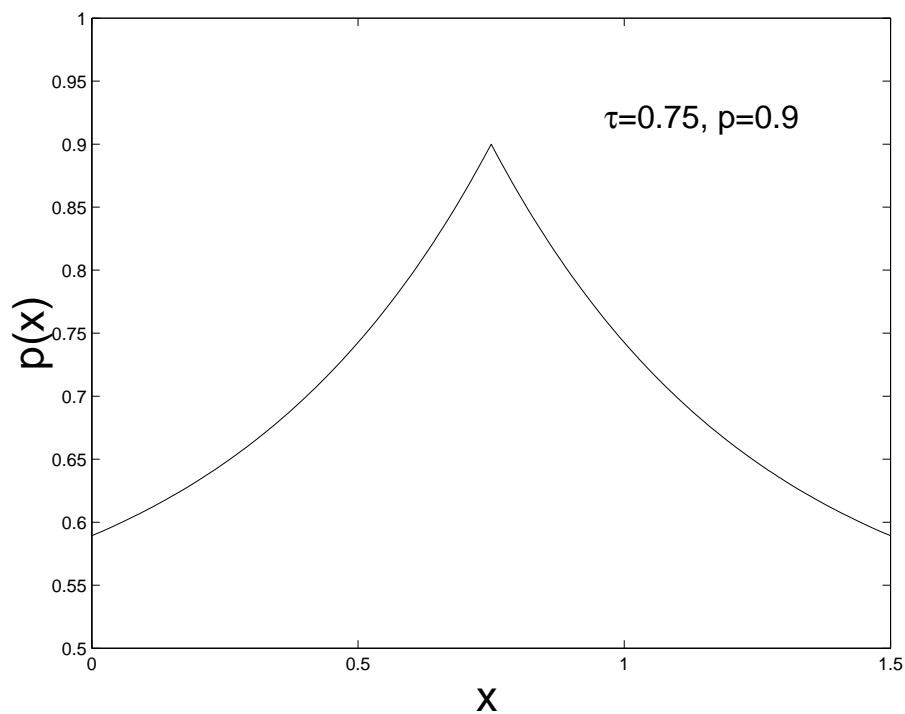
Nedärvning bredvid sjukdomsgen

Med en så kallad Poissonmodell för överkorsningar kan man visa att

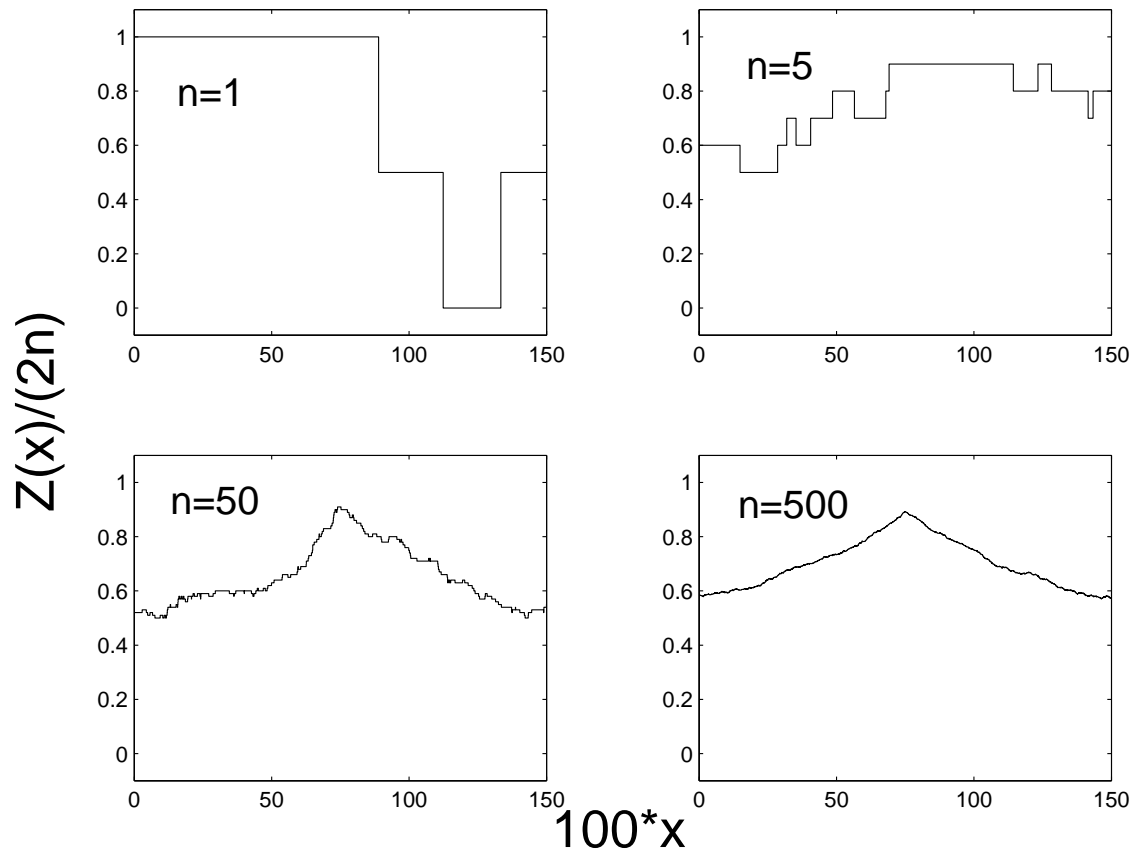
$$Z(x) \in \text{Bin}(2n, p(x)),$$

där

$$\begin{aligned} p(x) &= 0.5 + (p - 0.5)e^{-2|x-\tau|} \\ &= \text{proportionen DNA från samma mor-} \\ &\quad \text{farförälder, position } x. \end{aligned}$$

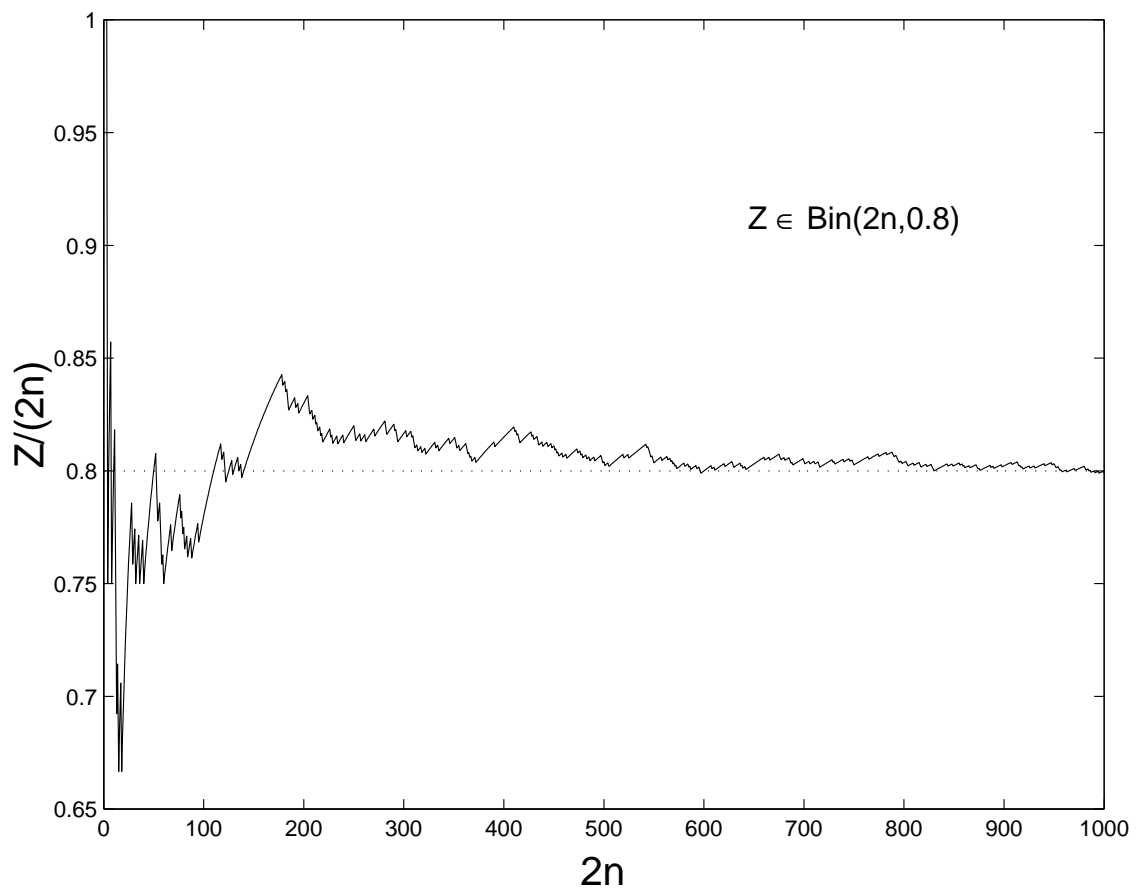


Plot av $Z(x)/(2n)$, $\tau = 0.75, p = 0.9$



Stora talens lag

$$\begin{aligned} Z(x) &\in \text{Bin}(2n, p(x)) \\ &\Downarrow \\ Z/(2n) &\rightarrow p(x) \text{ då } n \rightarrow \infty. \end{aligned}$$

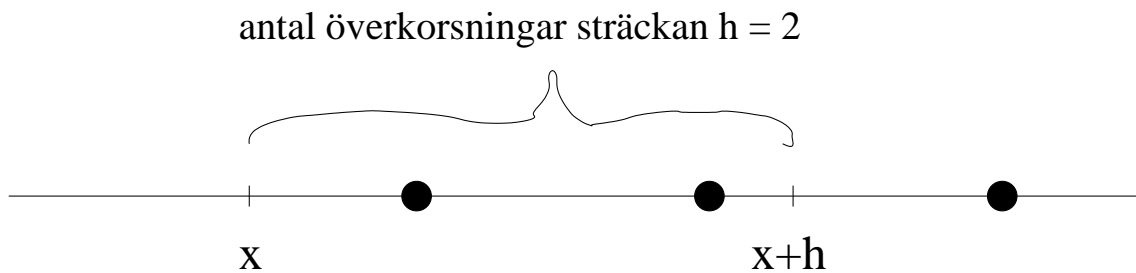


$$P(Z(x) = k) = \binom{2n}{k} (1 - p(x))^{2n-k} p(x)^k$$

Poissonmodell för överkorsningar

Anta överkorsningar uppträder

1. Oberoende
2. En i taget
3. Intensitet 1 överkorsning/ 10^8 baspar



Mät avstånd i enhet 10^8 baspar. Låt

$N_h =$ antal överkorsningar sträckan h .

1.-3. medför

$$P(N_h = k) = e^{-h} \frac{h^k}{k!}, \quad k = 0, 1, 2, \dots$$

som i sin tur ger formeln för $p(x)$ (se uppgifter).

Genletning

Intressant region för letning av τ :

$$\Omega = \{x; Z(x) \geq t\},$$

där tröskeln t bestäms så att

$$P(\tau \in \Omega) = 0.95.$$

$$\begin{aligned} L &= \text{genomsnittlig längd hos } \Omega \\ &\approx 3/(16(p - 0.5)^2 n). \end{aligned}$$

L	p	n
0.05	0.9	24
0.05	0.8	42
0.05	0.7	94
0.05	0.6	375
0.05	0.55	1 500
0.05	0.52	9 375

$L = 0.05$ svarar mot $5 \cdot 10^6$ baspar och igenomsnitt cirka 50 gener.

Generaliseringar

1. Större släktträd.
2. Mer generella sjukdomsmodeller
 - (a) Monogena (recessiv, dominant). Ex: Cystisk fibros, Huntingtons sjukdom.
 - (b) Polygena (flera gener bidrar till sjukdom)
 - (c) Heterogena (olika gener i olika populationer)
 - (d) Miljöeffekters samverkan med gener.
 - (e) Komplexa sjukdomar (polygena, heterogena, små p för varje gen, samverkan gen-miljö). Ex: Diabetes, Alzheimer, Parkinson, flera typer av ärftlig cancer (prostatacancer mm), hjärt-kärl sjukdomar.
3. Utnyttja överkorsningar 100-tals generationer tillbaks i tiden (associationsanalys) för finmappning.
4. Djurmodeller (större effekter, men svårare att dra slutsatser, kan planera experiment)

Mer läsning

Almgren, P., Bendahl, P-O. Bengtsson, H., Hössjer, O. and Perfekt, R. (2001). *Statistics in Genetics*. Lecture Notes, <http://www.maths.lth.se/matstat/kurser/statgen/>.

Hössjer, O. (2003). Assessing accuracy in linkage analysis by means of confidence regions. *Genetic Epidemiology*, **25**, 59-72.

Sham, P. (1998). *Statistics in Human Genetics*. Arnold applications of statistics, London.

Thomas, D.C. (2004). *Statistical Methods in Genetic Epidemiology*, Oxford Univesity Press, New York.