Uporaba statistike na področju kvantitativne genetike

Gregor Gorjanc

Univerza v Ljubljani, Biotehniška Fakulteta, Oddelek za zootehniko, Domžale (Rodica)

IBMI
Ljubljana, Slovenija
11. oktober 2010

UL, BF, Oddelek za zootehniko, Domžale (Rodica)

Izvleček

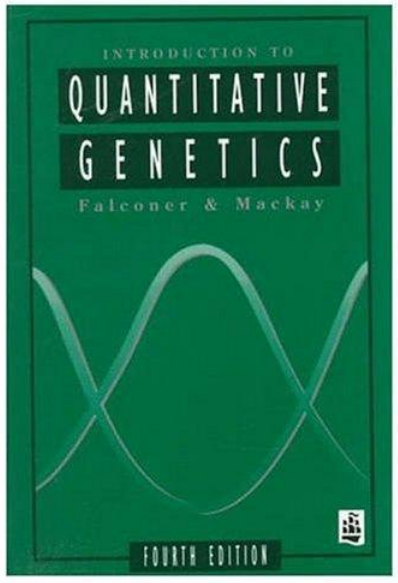
Kvantitativna genetika skuša odgovoriti na vprašanje kolikšen del fenotipske variabilnosti je povzročen z genetsko variabilnostjo.

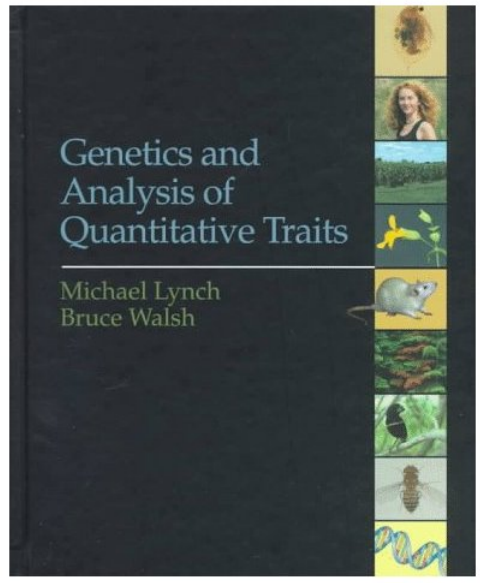
Vloga statističnih metod pri iskanju tega odgovora je ključna in razvoj nekaterih metod (npr. regresija, analiza variance, mešani model, model s pragovi, ...) je potekal hkrati z metodološkim razvojem na področju genetike kvantitativnih lastnosti.

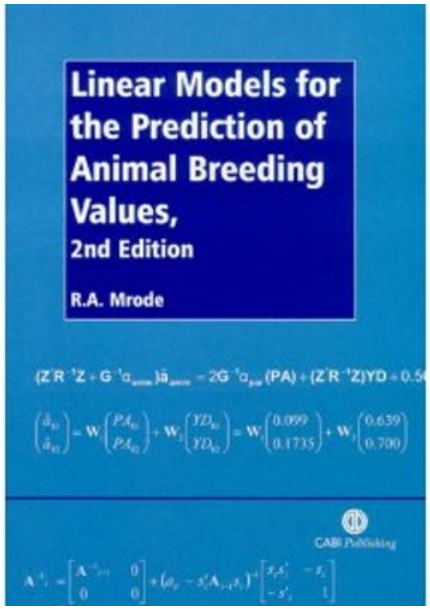
Tekom predavanja si bomo ogledali te metode s stališča statistike in kvantitativne genetike ter njihovo uporabo na podoročju selekcije domačih živali.

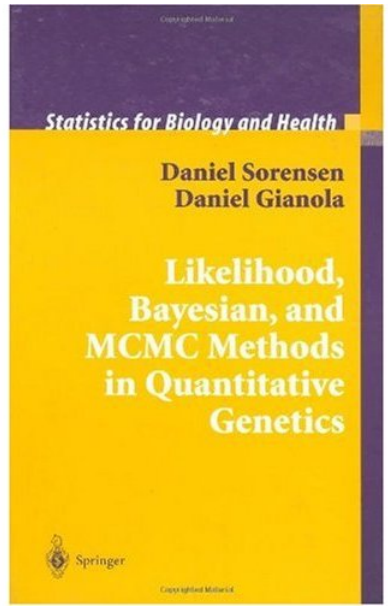
Kazalo

Nekaj literature

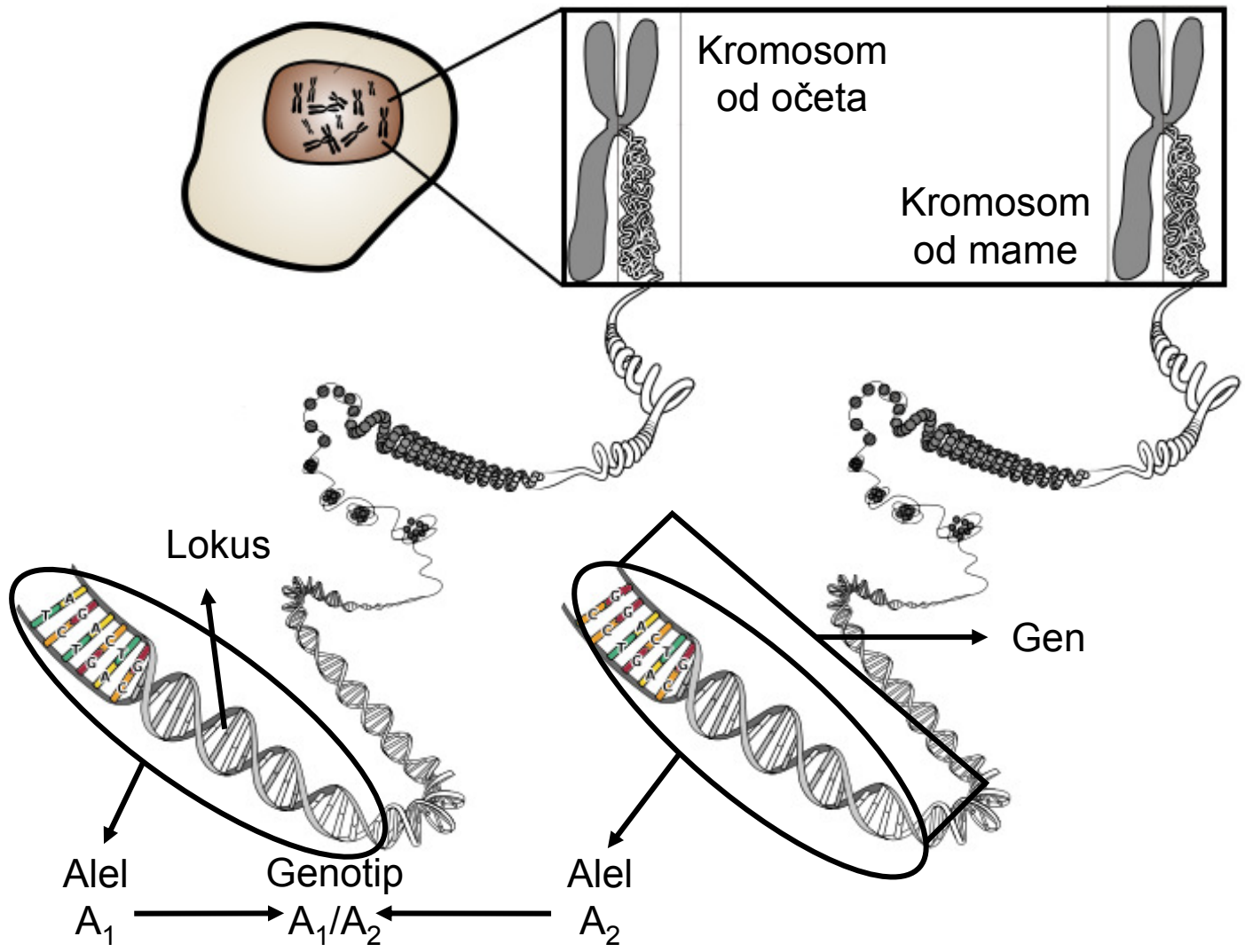








Genetika: DNA –> beljakovine –> fenotip



Področja genetike

- Populacijska genetika
- Kvantitatiivna genetika
- Evolucijska genetika
- Molekularna genetika
- Genetika
  - človeka
  - živali
  - rastlin
  - mikrobov
  - ...
- ...

Podobna “širina” in “prepredanje” kot pri področjih statistike

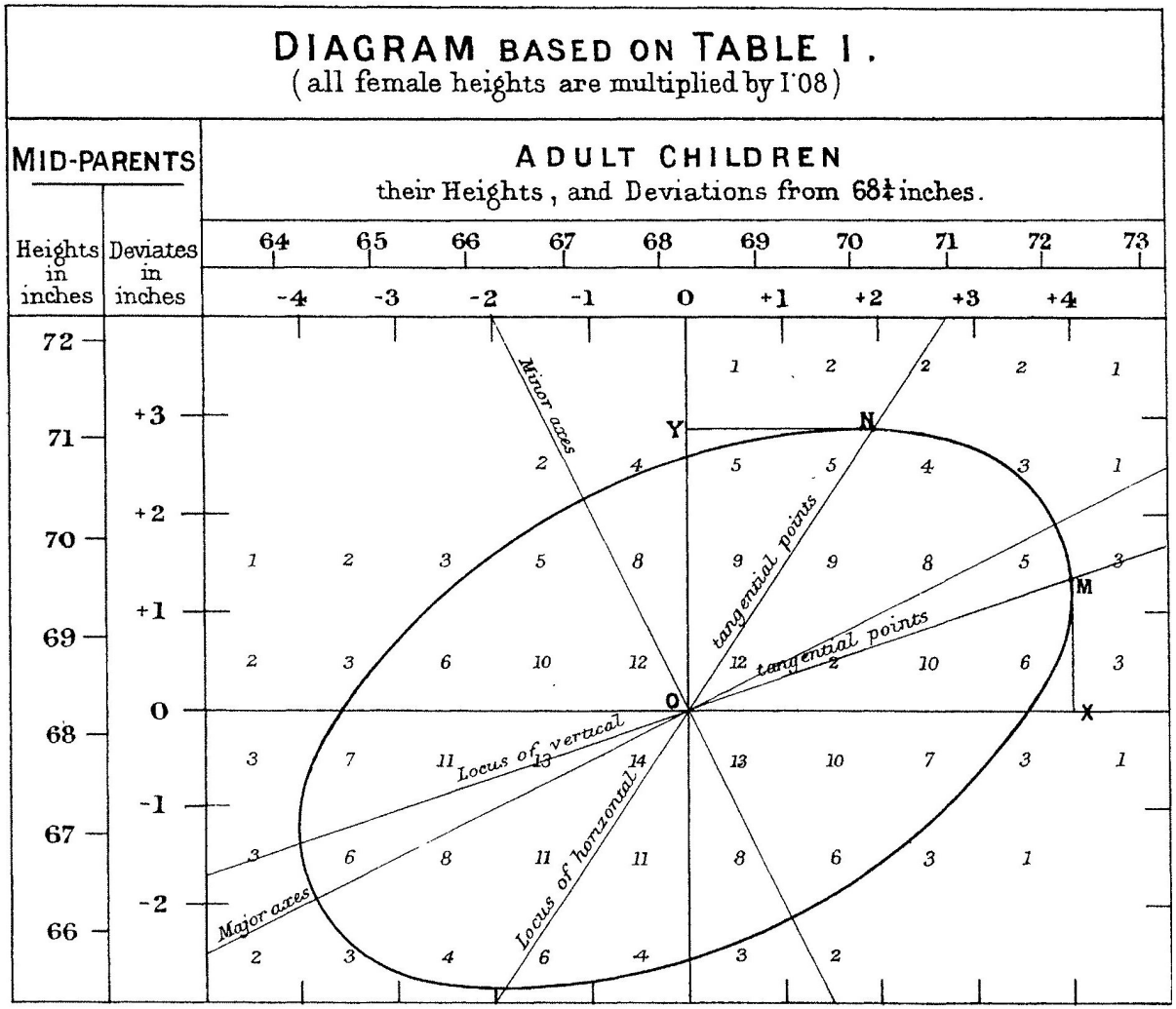
2. Začetki kvantitativne genetike

Začetki kvantitativne genetike

- Začetki genetike
  - študije vpliva genov na "enostavne" (diskretne) lastnosti (Mendel, Bateson, ...) –> **mendlisti**
  - študije "kompleksnih" (kvantitativnih) lastnosti (Pearson, Galton, Fisher, Wright, ...) –> **biometriki**
- **Galton** (1986) - regresija
- Nestrinjanje med mendlisti in biometriki ob koncu 19. stoletja
- **Fisher** (1918) - infinitezimalni model
- **Wright**, **Malecot** - inbriding in sorodstvo
- ...
- Razcvet, upad in ponovno razcvet z "genomsko revolucijo"

Galton (1886) - regresija

- Analiziral povprečno telesno višino staršev in njihovih (odraslih) otrok



Razprava: Zakaj je povezava (korelacija) pozitivna?

Regresija

- Bivariatna normalna porazdelitev fenotipskih vrednosti
  - telesna višina očeta - P_f
  - telesna višina sina - P_s

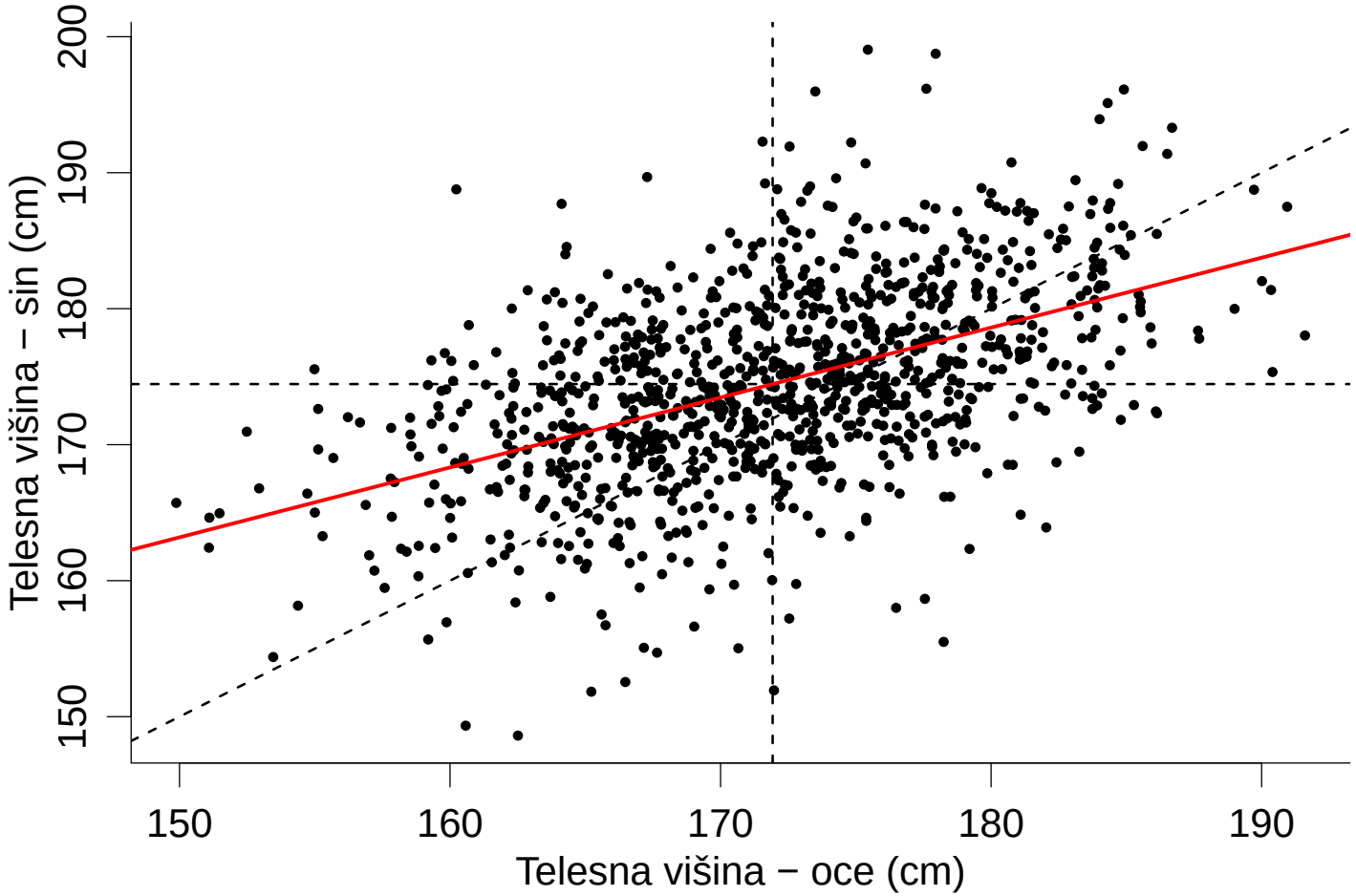
$$p\left(P_f, P_s\right) \sim N\left(\begin{array}{c} E\left(P_f\right) \\ E\left(P_s\right) \end{array}, \begin{array}{cc} Var\left(P_f\right) & Cov\left(P_f, P_s\right) \\ sym. & Var\left(P_s\right) \end{array}\right)$$

- Pogojno pričakovanje

$$E\left(P_s|P_f\right) = E\left(P_s\right) + Cov\left(P_s, P_f\right)\left(Var\left(P_f\right)\right)^{-1}\left(P_f - E\left(P_f\right)\right)$$

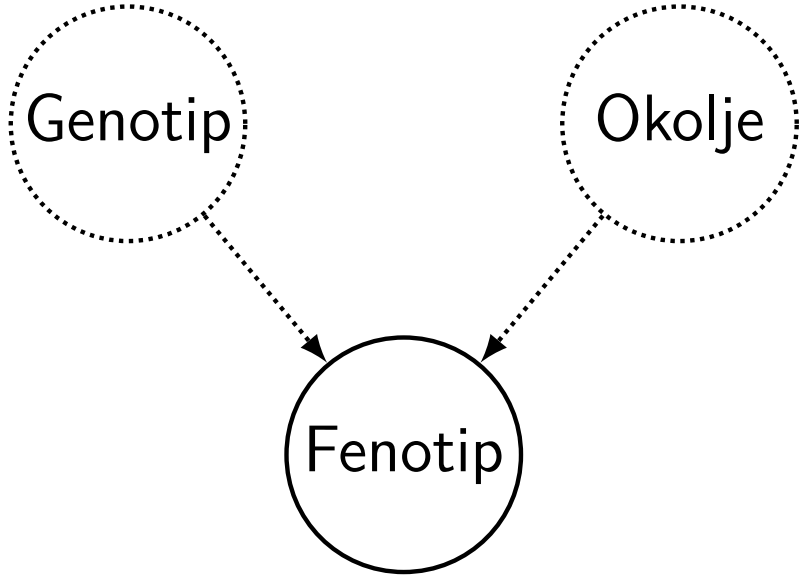
- Regresijski koeficient: $Cov\left(P_s, P_f\right)\left(Var\left(P_f\right)\right)^{-1}$

Regresija - Pearsonovi podatki



Razprava: Zakaj "regression/shrinkage" k povprečju?

Fisher (1918) - dekompozicija fenotipske vrednosti



- Koncept vrednosti

$$P = \mu + G + E$$

  - P - fenotipska vrednost - **lahko merimo!!!**
  - G - genotipska vrednost - **ne moremo meriti!!!**
  - E - deviacija zaradi okolja

Fisher (1918) - dekompozicija genotipske vrednosti

- Genotipska vrednost
  - En lokus: $G = A + D$
  - Več kot en lokus: $G = A + D + I$
  - A - additivna genotipska (plemenska) vrednost
    (učinek posameznih alelov)
  - D - deviacija zaradi dominance
    (interakcije med aleli na enem lokusu)
  - I - deviacija zaradi epistaze: $A \times A, A \times D, D \times D$
    (interakcije med aleli na različnih lokusih)

$$P = \mu + A + D + I + E$$

- Centralni limitni izrek
  - veliko število genov z majhnim učinkom + naključni vplivi iz okolja –> normalna porazdelitev

Fisher (1918) - analiza variance

- Analiza variance za **predpostavljeni** model

$$P \approx \mu + A + D + E$$

- Vzročne komponente fenotipske variance

$$\sigma_p^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2$$

- Dednostni delež (heritabiliteta)
  - v širšem smislu $h^2 = \sigma_g^2/\sigma_p^2$
  - v ožjem smislu $h^2 = \sigma_a^2/\sigma_p^2$
- Opazovane komponente variance - to lahko ocenimo iz zbranih podatkov
  - varianca med skupinami σ^2_{med}
  - varianca znotraj skupin $\sigma^2_{znotraj}$

Fisher (1918) - analiza variance II

- Analiza variance skupin sorodnikov –> opazovane komponente variance vsebujejo različne vzročne komponente variance

$$\begin{aligned} \sigma^2_{med} &= r_a\sigma^2_a + r_d\sigma^2_d + r_c\sigma^2_{ec} \\ \sigma^2_{znotraj} &= \sigma^2_P - \sigma^2_{med} \end{aligned}$$

  - r_a= verjetnost, da imata sorodnika enake alele
  - r_d= verjetnost, da imata sorodnika enak genotip
  - r_c = "verjetnost", da imata sorodnika "skupno okolje"

Skupine / σ^2_{med}	r_a	r_d	r_c
Enojajčni dvojčki	1	1	1
Dvojajčni dvojčki	$1/2$	$1/4$	1
Bratje in sestre (FSIB)	$1/2$	$1/4$	1
Pol-bratje in pol-sestre (HSIB)	$1/4$	0	0

Galton (1886) - revizija

- Predpostavljeni model za telesno višino sinov (P_s)

$$\begin{aligned} P_s &\approx \mu + A_s + E_s \\ &\approx \mu + {}^1\!/_2 A_f + {}^1\!/_2 A_m + E_s \end{aligned}$$

- Pogojno pričakovanje na podlagi telesne višine očetov (P_f)

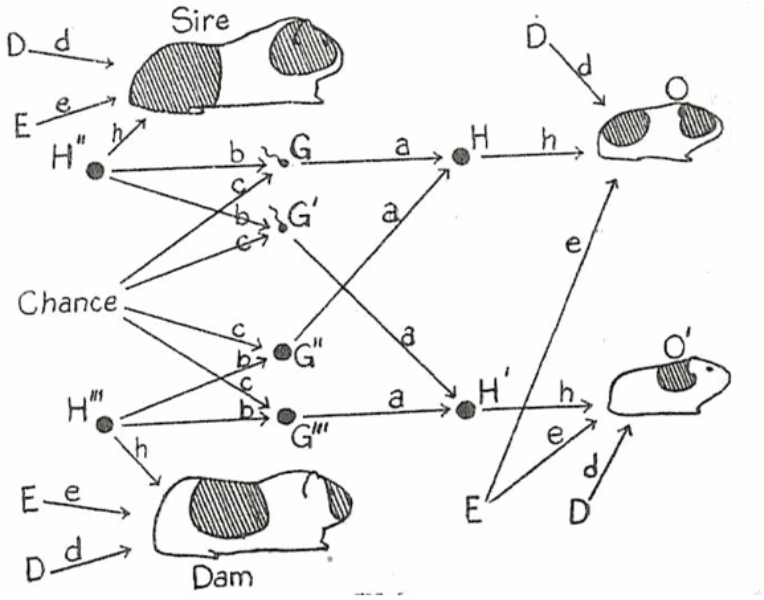
$$E\left(P_s|P_f\right) = E\left(P_s\right) + Cov\left(P_s, P_f\right)\left(Var\left(P_f\right)\right)^{-1}\left(P_f - E\left(P_f\right)\right)$$

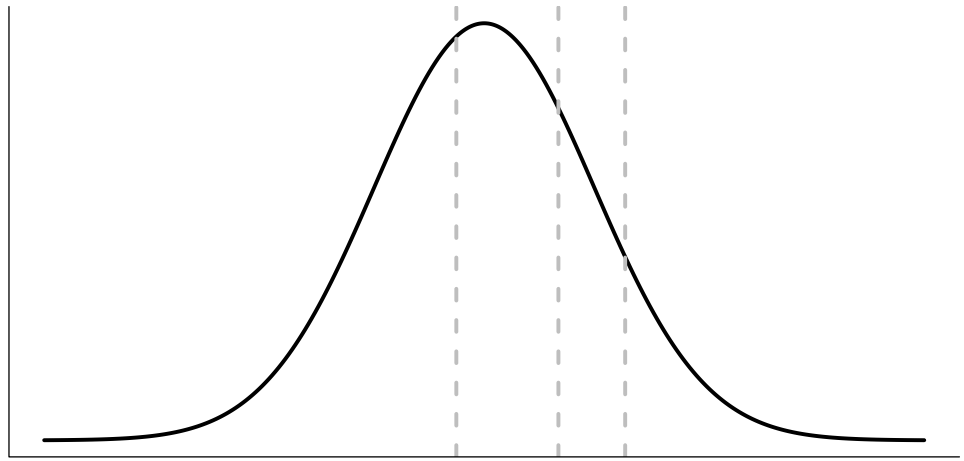
  - pričakovane vrednosti: $E\left(P_s\right) = E\left(P_f\right) = \mu$
  - varianca med očeti: $Var\left(P_f\right) = \sigma_p^2$
  - covarianca med sinovi in očeti: $Cov\left(P_s, P_f\right) = Cov\left({}^1\!/_2 A_f, A_f\right) = {}^1\!/_2 \sigma_a^2$
  - regresijski koeficient

$$Cov\left(P_s, P_f\right)\left(Var\left(P_f\right)\right)^{-1} = \frac{{}^1\!/_2 \sigma_a^2}{\sigma_p^2} = {}^1\!/_2 h^2 < 1$$

Wright

- S korelacijami definiral koncept sorodnosti na osnovi informacije iz rodovnikov:
  - **koef. inbridinga** - korelacija med gametami posameznika
  - **koef. sorodnosti** - korelacija med gametami posameznikov
- **Stezna metoda** (ang. path analysis) in **model s pragovi** (1934) (ang. threshold model) = probit model





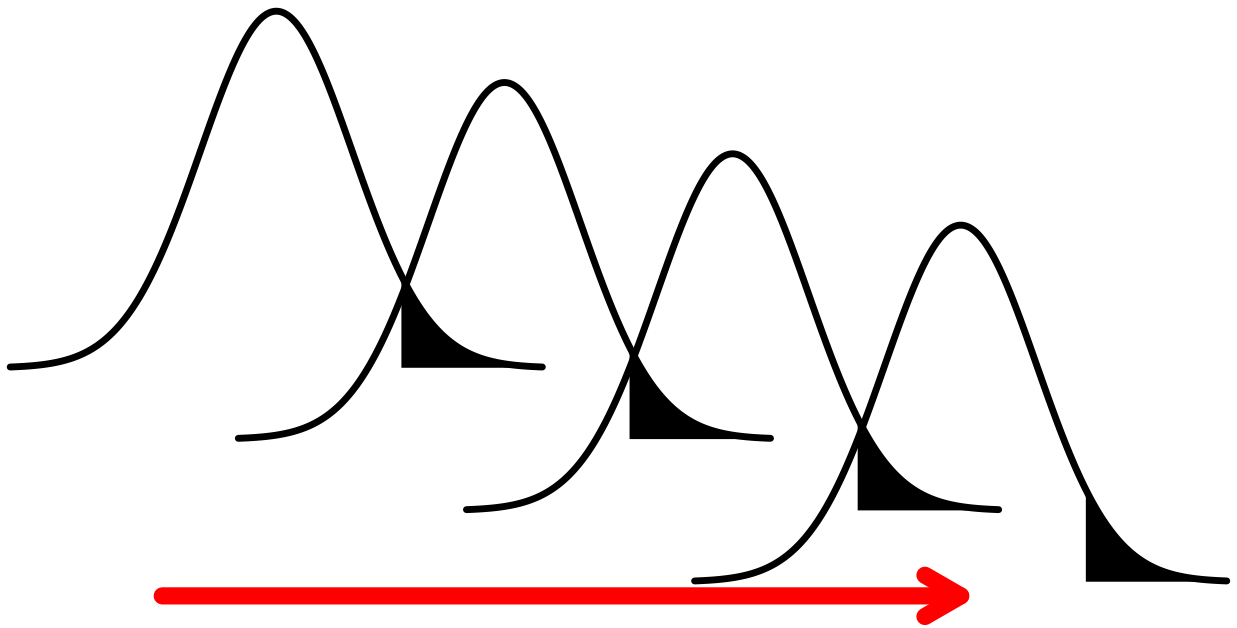
Raziskovalna "področja"

- **ljudje** –> biostatistika, genetika človeka (ang. human genetics)
  - cilj: splošno znanje in **zdravljenje**
- **rastline** –> genetika rastlin (ang. plant breeding and genetics)
  - cilj: –> splošno znanje in **selekcija** (=žlahtnjenje)
- **živali** –> genetika živali (ang. animal breeding and genetics)
  - cilj: splošno znanje in **selekcija**
  - "animal breeding and genetics"
    = živinoreja + genetika+ statistika + ...

3. Uporaba kvantitativne genetike pri selekciji domačih živali

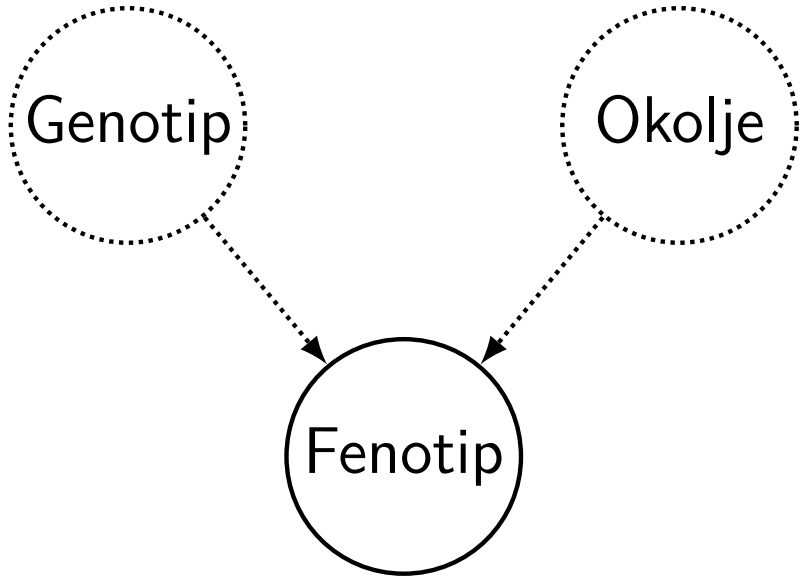
Selekcija

- Izmerimo **fenotipske vrednosti** kandidatov in izberemo (selekcionirami) tiste z najbolj zaželenimi vrednostmi (ang. **mass/phenotype selection**)
- Izbrani kandidati bodo starši naslednje generacije



Razprava: Ali se fenotipska vrednost prenaša s staršev na potomcev?

Dekompozicija fenotipske vrednosti



- **Genetsko vrednotenje** = statistično sklepanje o genotipski (plemenski) vrednosti posameznikov glede na zbrane **podatke** in **predpostavljeni model** (= **BLUP selection**)

Podatki?

- Individualni rejci in rejski programi
- Številne vrste (govedo, perutnina, prašiči, ovce, koze, konji, psi, mačke, ribe, raki, čebele, . . .)
- Številne (kompleksne) lastnosti - **fenotipske vrednosti**:
  - prireja (mleko, meso, jajca, . . .)
  - reprodukcija (št. potomcev, uspešnost pripusta, . . .)
  - konformacija telesa (višina vihra, obseg, . . .)
  - zdravje in dolgoživost
  - . . .
- **Rodovniki**
- Zadnja leta tudi podatki o genotipu živali
  - posamezni pomembni geni
  - veliko število DNA označevalcev

Henderson (1949+) - mešani model

- Mešani model - fiksni/sistematski (**b**) in naključni (**a**) vplivi

$$\begin{aligned} y_{ijk} &= \mu + b_i + a_j + e_{ijk} \\ \mathbf{y} &= \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \end{aligned}$$

- Predpostavke

$$\begin{aligned} p(\mathbf{a}) &\sim N(\mathbf{0}, \mathbf{G}), \quad \mathbf{G} = \mathbf{A}\sigma_a^2 \\ p(\mathbf{e}) &\sim N(\mathbf{0}, \mathbf{R}), \quad \mathbf{R} = \mathbf{I}\sigma_e^2 \end{aligned}$$

$$p\begin{pmatrix} \mathbf{y} \\ \mathbf{a} \\ \mathbf{e} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{Xb} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{matrix} \mathbf{ZGZ}^T + \mathbf{R} & \mathbf{ZG}^T & \mathbf{R} \\ & \mathbf{G} & \mathbf{0} \\ \textit{sym.} & & \mathbf{R} \end{matrix} \right)$$

A - matrika sorodstva (Wright)

Mešani model - poimenovanje

- Henderson: model živali, ker je modeliral prirejo živali
  - model očeta - upoštevamo rodovnik samo med očeti
  - model očeta in mame - upoštevamo rodovnik samo med očeti
  - ...
  - bolj splošno: **mešani model z rodovniki**
- Mešani model (ang. mixed model)
- Hierarhični model (ang. hierarchical model)
- Večnivojski??? model (ang. multilevel model)
- ...

Henderson (1949+) - sistem enačb

- Metoda najmanjših kvadratov

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} \right) \left(\widehat{\mathbf{b}} \right) = \left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \right)$$

- Mešani model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e}$$

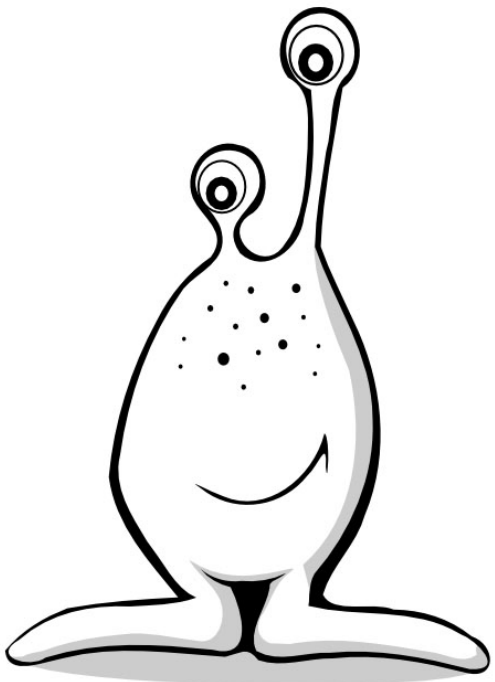
$$\widehat{\mathbf{a}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}}\right)$$

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

$$\mathbf{G}^{-1} = \mathbf{A}^{-1}\sigma_a^{-2} \qquad \mathbf{R}^{-1} = \mathbf{I}\sigma_e^{-2}$$

- $\widehat{\mathbf{b}}$ - najboljša linearna nepristranska cenilka (**BLUE**)
- $\widehat{\mathbf{a}}$ - najboljša linearna nepristranska napoved??? (**BLUP**)

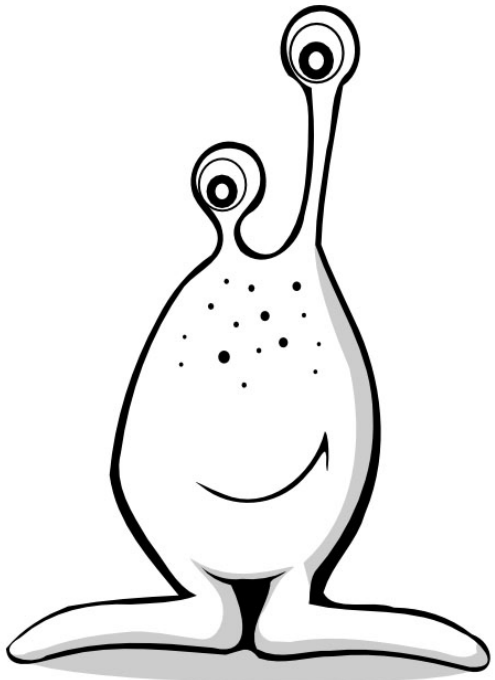
Primer “vesoljcev” - podatki



Posameznik	Oče	Mama	Skupina	Fenotip
1	/	/	/	/
2	/	/	1	103, 106
3	2	1	1	98
4	2	/	2	101
5	4	3	2	106
6	2	3	2	93
7	5	6	/	/
8	5	6	/	/
9	/	/	/	/
10	8	9	1	109

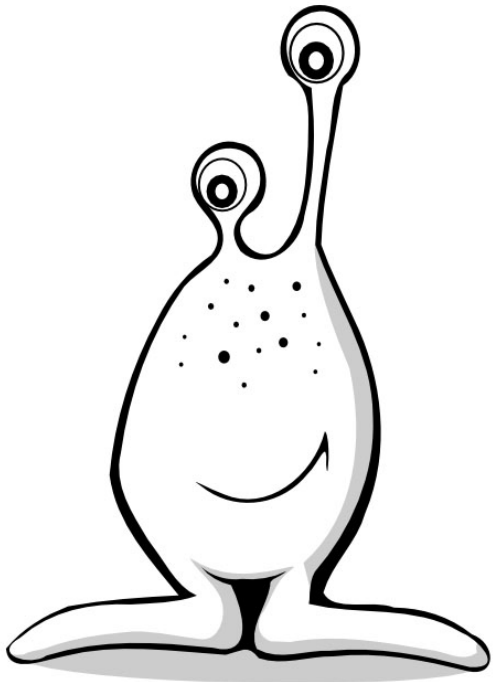
Slika: Jouke

Primer “vesoljcev” - grafični model

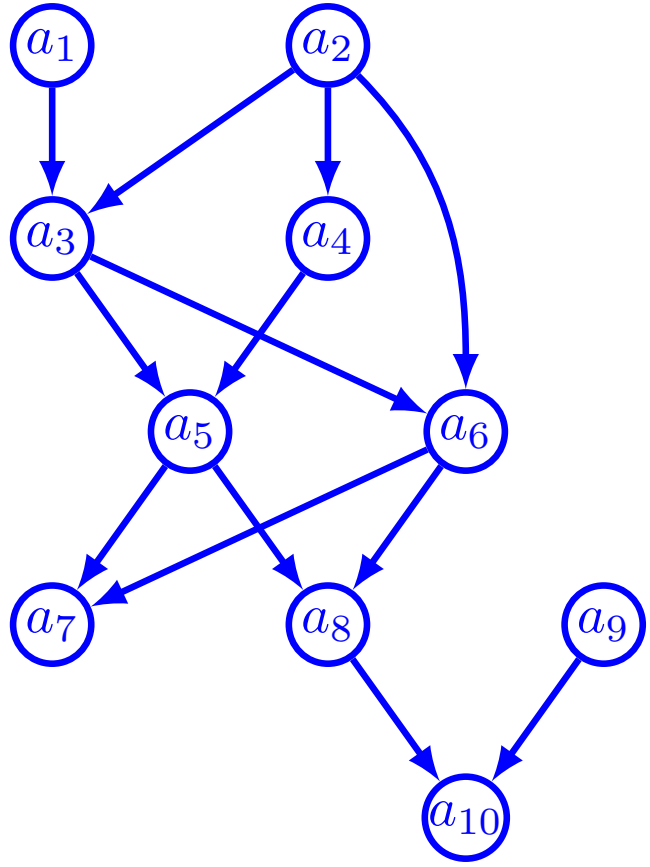


Slika: Jouke

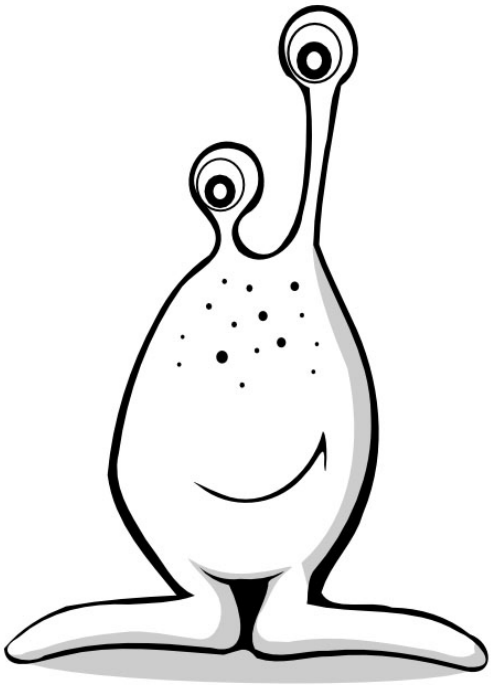
Primer “vesoljcev” - grafični model



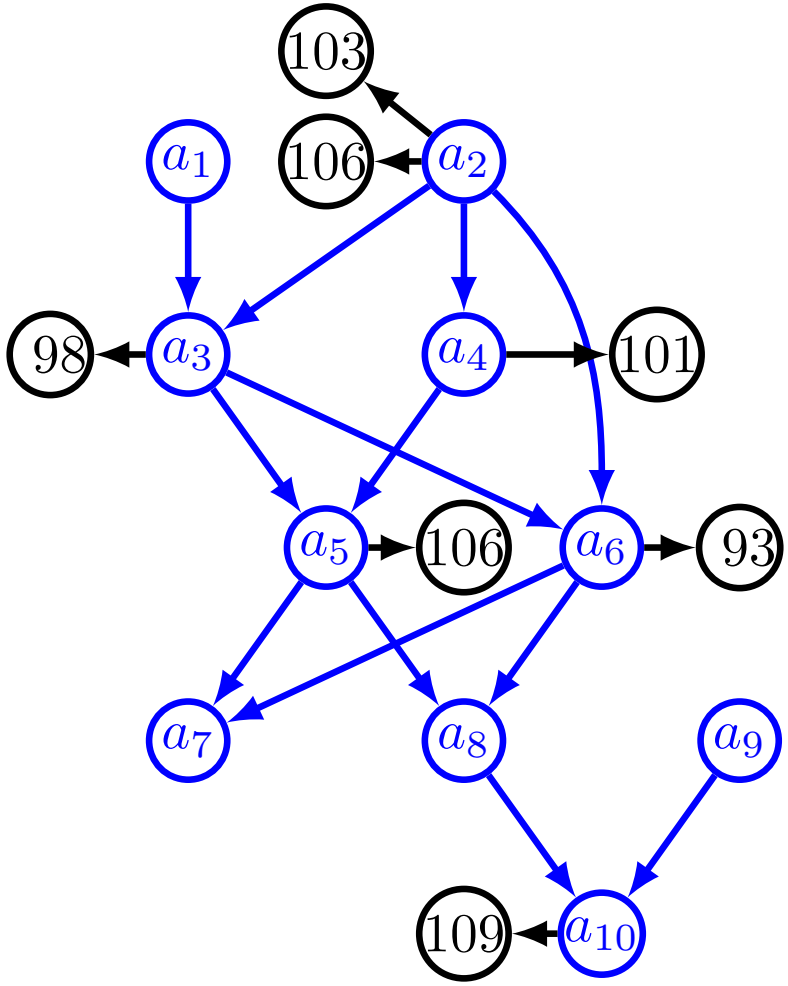
Slika: Jouke



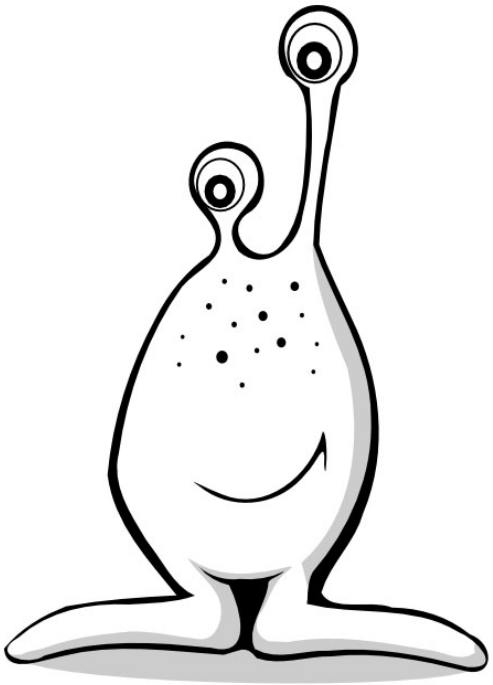
Primer “vesoljcev” - grafični model

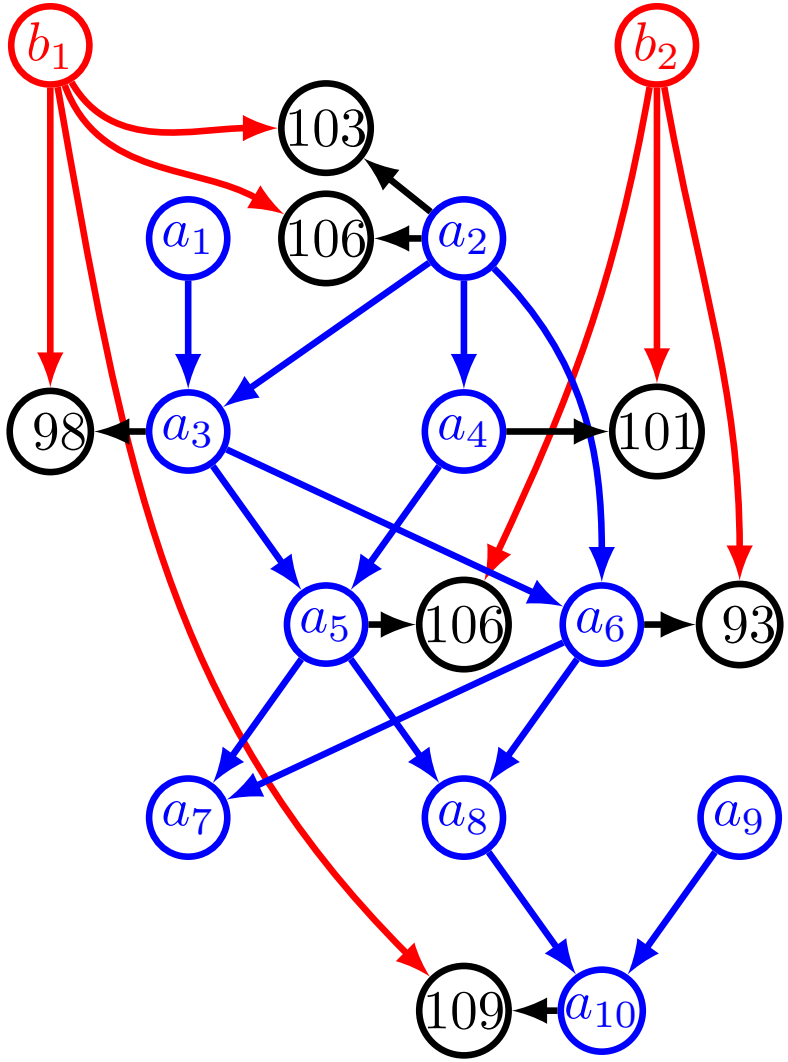


Slika: Jouke



Primer “vesoljcev” - grafični model





Slika: Jouke

Primer “vesoljcev” - R

Demonstracija v R-ju

Programje

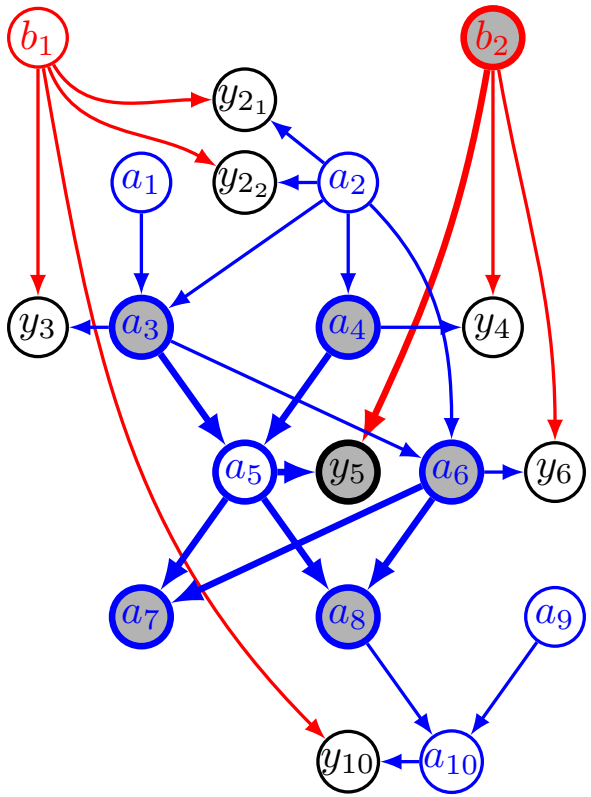
- Aplikacije lahko zajemajo tudi več 100 tisoč ali milijon živali in toliko ali še več fenotipskih vrednosti
  - Primer ~150.000 telitev = fenotipskih vrednosti, ~75.000 krav + ~1.000 bikov
- Specializirani programi (rešitev sistema enačb in/ali ocena komponent variance)
  - ASREML
  - BLUPf90
  - DMU
  - PEST
  - SurvivalKit
  - VCE
  - WOMBAT
  - ...

Posebnosti

- Statistični pomen inverze matrike sorodstva $\mathbf{A}^{-1}$
- Bayesovski pogled
- Komponente variance?

Statistični pomen inverze matrike sorodstva $\mathbf{A}^{-1}$

- Pogojna neodvisnost spremenljivk
  - $\mathbf{A}^{-1}_{i,j} = 0$ –> pogojna neodvisnost
  - $\mathbf{A}^{-1}_{i,j} \neq 0$ –> pogojna odvisnost
  - grafična predstavitev modela = grafični modeli
- Aditivna genotipska vrednost posameznika
  = f(povp. staršev, fenotipsko odstopanje, povp. potomcev)



Bayesovski pogled

- Model za analizirano spremenljivko

$$p\left(\mathbf{y}|\mathbf{b},\mathbf{a},\mathbf{R}\right) \sim N\left(\mathbf{Xb}+\mathbf{Za},\mathbf{R}\right)$$

- Parametri: $\mathbf{b}$, $\mathbf{a}$, σ_a^2, σ_e^2 –> apriorne porazdelitve

$$\begin{aligned} p\left(\mathbf{b}|\boldsymbol{\mu_b},\mathbf{B}\right) &\sim N\left(\boldsymbol{\mu_b},\mathbf{B}\right) \\ p\left(\mathbf{a}|\boldsymbol{\mu_a},\mathbf{G}\right) &\sim N\left(\boldsymbol{\mu_a},\mathbf{G}\right) \\ p\left(\sigma_a^2|\ldots\right), p\left(\sigma_a^2|\ldots\right) &\sim \ldots \end{aligned}$$

- Posteriorna porazdelitev za $\mathbf{b}$ in $\mathbf{a}$ če so variance poznane

$$p\left(\mathbf{b},\mathbf{a}|\mathbf{R},\boldsymbol{\mu_b},\mathbf{B},\boldsymbol{\mu_a},\mathbf{G}\right) \propto p\left(\mathbf{y}|\mathbf{b},\mathbf{a},\mathbf{R}\right) p\left(\mathbf{b}|\boldsymbol{\mu_b},\mathbf{B}\right) p\left(\mathbf{a}|\boldsymbol{\mu_a},\mathbf{G}\right)$$

Bayesovski pogled II

$$
\begin{aligned}
p\left(\mathbf{b}, \mathbf{a} | \mathbf{R}, \boldsymbol{\mu_b}, \mathbf{B}, \boldsymbol{\mu_a}, \mathbf{G}\right) &\sim N\left(\widehat{\theta}, \mathbf{C}^{-1}\sigma_e^2\right) \\
\mathbf{C}\widehat{\theta} &= \mathbf{r} \\
\widehat{\theta} &= \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \end{pmatrix} \\
\mathbf{C} &= \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} + \mathbf{B}^{-1} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \\
\mathbf{r} &= \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} + \mathbf{B}^{-1}\boldsymbol{\mu_b} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} + \mathbf{G}^{-1}\boldsymbol{\mu_a} \end{pmatrix}
\end{aligned}
$$

Ocena komponent variance

- Metoda največjega verjetja
  - Maximum Likelihood (ML)

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})}{n}$$

  - Restricted/Residual Maximum Likelihood (REML; Patterson & Thompson, 1971)

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})}{n - p}$$

    - EM, AI-REML, ...
- Bayesovski pristopi

Vprašanja?