Heritabilita III

- metody odhadu koeficientu heritability

prof. Ing. Tomáš Urban, Ph.D.
urban@mendelu.cz

ANOVA

Princip:

Detekce důležitých rozdílných zdrojů efektů

Určit jejich příspěvek na celkové varianci

Variance je odvozena ze součtu čtverců a stupňů volnosti

Nutné jedince ve skupinách se stejným stupněm příbuznosti

Skupiny polosourozenců podle otce

Rodiče – potomci

Kovariance mezi členy rodin nebo skupin = komponenta variance mezi skupinami

Rozčlenění součtu čtverců (SS) podle zdrojů variance (skupina zvířat) a výpočet středního čtverce (MS) ~ variance

Sire model – 1 f ANOVA

- Odhad korelace polosourozenců

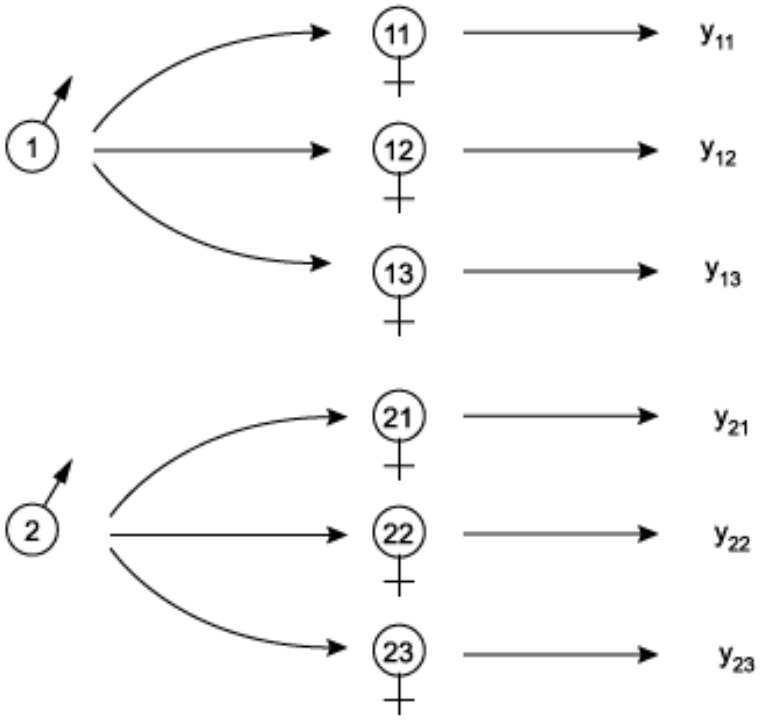
$$y_{ij} = \mu + a_i + e_{ij}$$

$$V_S = \sigma_S^2 = \frac{1}{4}\sigma_A^2$$

$$V_e = \sigma_e^2 = \frac{3}{4}\sigma_A^2 + \sigma_E^2 \qquad \sigma_y^2 = \sigma_S^2 + \sigma_e^2$$

- předpoklad, že otcové a matky jsou nepříbuzní, náhodně páření, bez selekce
- balancovaný design: *p* otců (sire) pářeno s *n* matkami (dam) ⇒ 1 potomka

Design pokusu pro Sire model



Variance mezi skupinami polosourozenců = kovarianci mezi polosourozenci ve skupině

$$cov_{(polos.)} = cov(y_{ij}, y_{ik}) = \sigma^2_S$$

$$4\sigma_S^2 = \sigma_A^2$$

To lze pomocí ANOVA odhadnout

Sire model – tabulka ANOVA

Zdroj proměnlivosti	df	SS	MS	E(MS)
Mezi rodinami (mezi otci)	**p – 1**	$SS_S = \sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2$	$MS_S = \frac{SS_S}{(p-1)}$	$\sigma_e^2 + n_0\sigma_g^2$
V rodinách (reziduální)	**n – p**	$SS_e = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$MS_e = \frac{SS_e}{(n-p)}$	σ_e^2
Celkem	**n – 1**	$SS_c = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	$MS_c = \frac{SS_c}{(n-1)}$	

Vážený počet potomků na 1 otce

$$n_0 = \frac{n - \left(\sum n_i^2 / n\right)}{n-1}$$

$$MS_e \doteq \sigma_e^2$$

$$MS_a = \sigma_e^2 + n_0\sigma_g^2 = MS_e + n_0\sigma_g^2$$

$$\sigma_g^2 = \frac{MS_a - MS_e}{n_0}$$

Intraklasní korelační koeficient

$$\rho = r_i = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

$$h^2 = 4\rho = 4\frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = 4\frac{\sigma_g^2}{\sigma_P^2}$$

Vybalancovaná data

$$se_{h^2} = 4.s_\rho = 4.\sqrt{\frac{2.(1-\rho)^2(1+(n_0-1)\rho)^2}{n_0(n_0-1)(p-1)}}$$

Nevybalancovaná data

$$se_{h^2} = 4.s_\rho = 4.\sqrt{\frac{2.(n-1)(1-\rho)^2(1+(n_0-1)\rho)^2}{n_0(n_0-p)(p-1)}}$$

Závěr výpočtu

- odhad koeficientu dědivosti
- odhad střední chyby h^2
- intervalu spolehlivosti (hranice platnosti, např. 95 %)

$$h^2 \pm se_{h^2} \qquad \ldots \leq h^2 \leq \ldots$$

Př. 1 faktorové ANOVA pro výpočet h^2

-skupin polosourozenců

Statistický model jednofaktorové analýzy variance:

$$y_{ij} = \mu + a_i + e_{ij}$$

y_{ij} – užitkovost j-tého potomka po i-tém otci

μ – obecný průměr populace

a_i – vliv i-tého otce

e_{ij} – ostatní nahodilé vlivy

Výpočet součtu čtverců odchylek od průměru:

- mezi otci $$SS_a = \sum_{i=1}^{p} \frac{Y_{i\bullet}^2}{n_i} - \frac{Y_{\bullet\bullet}^2}{n}$$
- uvnitř skupin podle otců (reziduální)

$$SS_e = \sum_{i=1}^{p} \sum_{j=1}^{m_j} y_{ij}^2 - \sum_{i=1}^{p} \frac{Y_{i\bullet}^2}{n_i}$$

n	O_1	O_2	O_3	O_4	O_5
1	717	732	603	648	690
2	704	694	731	669	650
3	753	691	737	693	788
4	700	631	678	718	678
5	675	683	747	606	611
6	793	592	763	669	674
7	691	680	687	657	658
8	687	618	618	600	717
Σ	5720	5321	5564	5260	5466

skupina	$Y_{i\bullet}$	$Y_{i\bullet}^2$	$Y_{i\bullet}^2/n_i$	$\sum y_{ij}^2$
O_1	5720	32718400	4089800,00	4100638
O_2	5321	28313041	3539130,13	3554379
O_3	5564	30958096	3869762,00	3894894
O_4	5260	27667600	3458450,00	3469684
O_5	5466	29877156	3734644,50	3753878
	$Y_{\bullet\bullet}$ = 27331		18691786,63	**18773473**
	$Y_{\bullet\bullet}^2$ = 746983561		$\sum Y_{i\bullet}^2/n_i$	$\sum\sum y_{ij}^2$

p = 5 n = 40 $n_i = n_0$ = 8

Výsledek analýzy variance z programu SAS

Dependent Variable: potomek

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (a)	4	17197.60000	**4299.40000**	1.84	0.1428
Error (e)	35	81686.37500	**2333.89643**		
Corrected Total	39	98883.97500			

$MS_a = \sigma_e^2 + n_0\sigma_g^2$

$MS_e = \sigma_e^2$

Výpočet odhadu **genetické variance** podle otců:

$$MS_a = \sigma_e^2 + n_0\sigma_g^2 = MS_e + n_0\sigma_g^2$$

$$\sigma_g^2 = \frac{MS_a - MS_e}{n_0} = \frac{4299{,}4 - 2333{,}89643}{8} = 245{,}683$$

$$\rho = r_i = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = \frac{245{,}368}{245{,}368 + 2333{,}896} = 0{,}0952$$

$$h^2 = 4\rho = 4\frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = 4\frac{\sigma_g^2}{\sigma_P^2} \qquad se_{h^2} = 4.s_\rho = 4.\sqrt{\frac{2.(1-\rho)^2\left(1+(n_0-1)\rho\right)^2}{n_0(n_0-1)(p-1)}}$$

$$h^2 \pm se_{h^2} = 0{,}38 \pm 0{,}57$$

2 faktorová hierarchická ANOVA

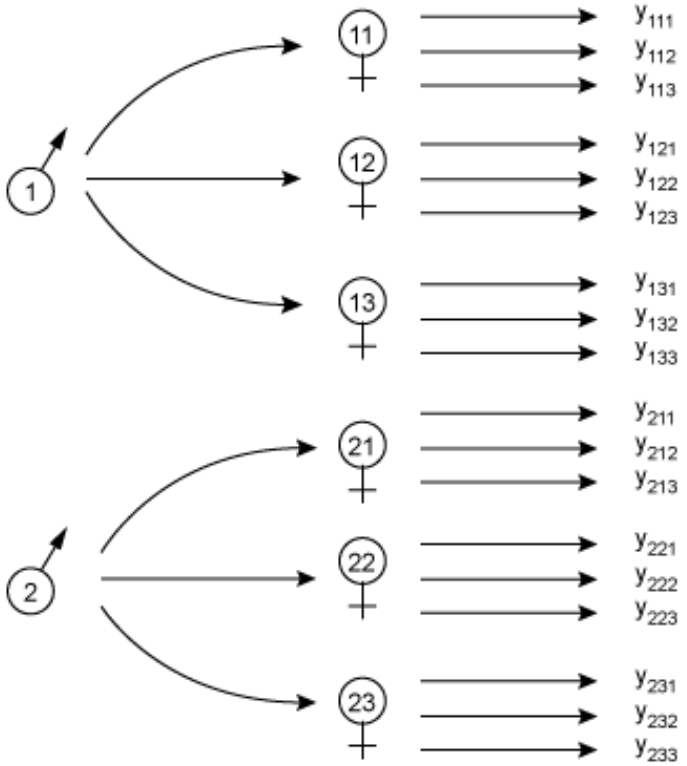
- Odhad korelace u **vlastních sourozenců** a **polosourozenců**
- Stanovení komponent variance mezi a v rodinách vlastních sourozenců

$$\mathbf{y_{ijk} = \mu + a_i + b_{ij} + e_{ijk}}$$

$$\sigma_y^2 = \sigma_s^2 + \sigma_d^2 + \sigma_e^2$$

- předpoklad, nejsou efekty dominance a společného prostředí
- balancovaný design: *p* otců (sire) pářeno s *m* matkami (dam) $\Rightarrow$ *n* potomky

Design pokusu pro analýzu úplných sourozenců a polosourozenců



Variance mezi skupinami úplných sourozenců = kovarianci mezi úplnými sourozenci ve skupině

Zdroj proměnlivost	df	SS	MS	E(MS)
Mezi otci (mezi rodinami)	**p – 1**	$SS_S = \sum_{i=1}^{p}\sum_{j=1}^{m_i} n_{ij}(\bar{y}_i - \bar{y})^2$	$MS_S = \frac{SS_S}{(p-1)}$	$\sigma_e^2 + k_2\sigma_{g_M}^2 + k_3\sigma_{g_O}^2$
Mezi matkami (uvnitř otců)	**m – p**	$SS_d = \sum_{i=1}^{p}\sum_{j=1}^{m_i} n_{ij}(\bar{y}_{ij} - \bar{y}_i)^2$	$MS_d = \frac{SS_d}{(m-p)}$	$\sigma_e^2 + k_1\sigma_{g_M}^2$
Mezi potomky (v otcích a matkách)	**n – m**	$SS_e = \sum_{i=1}^{p}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij})^2$	$MS_e = \frac{SS_e}{(n-m)}$	σ_e^2
Cekem	**n – 1**	$SS_c = \sum_{i=1}^{p}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y})^2$	$MS_c = \frac{SS_c}{(n-1)}$	

Odhad h^2 u vlastních sourozenců a polosourozenců

$$\sigma^2_{g_M} = \frac{MS_b - MS_e}{k_1}$$

$$\sigma^2_{g_O} = \frac{MS_a - MS_e - k_2\sigma^2_{g_M}}{k_3}$$

Když $k_1 = k_2$: $\sigma^2_{g_O} = \frac{MS_a - MS_b}{k_3}$

potomků/matku = matek/otce

a) podle otců

$$h^2_O = 4\rho_O = 4\frac{\sigma^2_{g_O}}{\sigma^2_{g_O} + \sigma^2_e} = 4\frac{\sigma^2_{g_O}}{\sigma^2_P}$$

b) podle matek

$$h^2_M = 4\rho_M = 4\frac{\sigma^2_{g_M}}{\sigma^2_{g_M} + \sigma^2_e} = 4\frac{\sigma^2_{g_M}}{\sigma^2_P}$$

c) podle otců a matek

$$h^2_{O+M} = 2\rho_{O+M} = 2\frac{\sigma^2_{g_O} + \sigma^2_{g_M}}{\sigma^2_{g_O} + \sigma^2_{g_M} + \sigma^2_e} = 2\frac{\sigma^2_{g_O} + \sigma^2_{g_M}}{\sigma^2_P}$$

ANOVA v maticovém zápisu

Lineární model ANOVA

y = μ + efektA + efektB + … + e

můžeme vyjádřit v maticích:

y = Xb + e

X je matice designová s 0 a 1, které sledují experimentální plán a jeho lineární model

Zobecněný lineární model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

y je sloupcový vektor vlastností/proměnné pro N jedinců	**X** je designová matice (N × r)
b je vektor parametrů	**e** je vektor reziduí

Designová matice X

Jedinec	Otec
1	1
2	1
3	3
4	2
5	3
6	1
7	2
8	1
9	3
10	2

X

		Otec		
		O1	O2	O3
Jedinec	1	1	0	0
	2	1	0	0
	3	0	0	1
	4	0	1	0
	5	0	0	1
	6	1	0	0
	7	0	1	0
	8	1	0	0
	9	0	0	1
	10	0	1	0

Řešení odhadů nejmenších čtverců vektoru **b**

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

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Metody založené na ML

Maximum Likelihood (ML)

REstricted Maximum Likelihood (REML)

Maximilizuje pravděpodobnost pozorovaných dat daných parametrů

Nebalancovaná data

Komplexní rodokmenová struktura (matice příbuznosti)

Simultánní korekce pro fixní efekty

Vyžaduje známou distribuci (normální)

Odhady jsou nevychýlené a jsou vždy v parametrovém prostoru

Funkce hustoty pravděpodobnosti normálního rozdělení:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

Očekávané průměry E(y) = Xb a var(y) = V

Logaritmus věrohodnostní funkce:

$$L(b, V \mid X, y) = -\tfrac{1}{2} N \log(2\pi) - \tfrac{1}{2}\log(|V|) - \tfrac{1}{2}(y - Xb)' V^{-1} (y - Xb)$$

- Rovnice dává pravděpodobnost parametrů (*b*, *V*) daných dat (*X*, *y*)
- Na pravé straně
  - první dva výrazy jsou očekávané hodnoty
  - poslední výraz je součet čtverců

První derivace: $\delta(\log L)/\delta \mathbf{b} = -2\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})$

Derivace = 0 $\hat{\mathbf{b}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$

Stejné jako pro LS odhady

Příklad algoritmu REML

1 Řešení rovnic smíšeného modelu s a priory hodnotou komponent variance (poměr)

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'Y \end{bmatrix}$$

2 Řešení komponent variance z MME

$$\sigma_a^2 = \left[\hat{a}' A^{-1} a + tr(A^{-1}C)\sigma_e^2 \right] / q$$

$$\sigma_e^2 = \left[y'y - \hat{b}'X'y - \hat{a}'Z'y \right] / (N - r(X))$$

Nové λ ($= \sigma_e^2/\sigma_a^2$) a iterovat mezi 1 a 2

Proč je REML lepší než ANOVA?

Je přesnější

Používá rovnice smíšeného modelu, takže využívá příbuzenské vztahy všech zvířat (animal model)

Má tedy vlastnosti jako BLUP

Dovoluje řešit více komplikované smíšené modely (maternální efekty, multiple traits ...) jako BLUP

ALE při vybalancovaném pokusu jsou výsledky odhadů REML a ANOVA stejné

Heritability Estimates of Protein %, Fat %, Lactose %, Non Fat Solids and Total Solids of Dairy Cattle in Northern Thailand

N. Chongkasikita, T. Vearasilpa and U. ter Meulenb

Deutscher Tropentag 2002, Witzenhausen, October 9-11, 2002, Conference on International Agricultural Research for Development

530 krav, 3 chovy

protein %, tuk %, laktóza %, sušina bez tuku a celková sušina

Pevné efekty: stádo-rok, sezóna, podíl HF skotu, počet dní laktace (regrese)

AM BLUP, použití **REML** programem VCE4 (Groeneveld, 1998).

$$y_{ijklm} = \mu + C_i + HF_j + HY_k + S_l + \mathbf{A_m} + b(X_{ijklm} - X) + E_{ijklm}$$

y_{ijklm}	Produkční vlastnosti
μ	průměr
C_i	Skupiny 1-10 podle % oblasti bíle zbarvené srsti (barva)
HF_j	Skupiny 1-5 podle % Holstein Friesian plemene u krav
HY_k	Stádo - Rok (1997, 1998, 1999, 2000 a 2001)
S_l	Období otelení (zima, léto a deště)
A_m	Jedinci (zvířata)
$b(X_{ijklm}-X)$	Věk při prvním otelení jako kovariata
E_{ijklm}	Náhodné reziduální efekty

	protein %	tuk %	laktóza %	sušina bez tuku	celková sušina
Heritabilita	**0,342**	**0,379**	**0,238**	**0,260**	**0,133**
V_A	0,041	0,130	0,022	0,963	0,036
V_E	0,079	0,212	0,069	2,736	0,238

Odhady komponent variance

Proces rozčlenění fenotypové variance na její komponenty (V_A a V_E)

Proč odhadujeme komponenty variance?

- Lepší porozumění mechanizmu kontrolující vlastnost
- Nutné pro predikci plemenných hodnot
- Nutné pro optimalizaci šlechtitelských programů

Měly by být komponenty variance znovu odhadovány v čase?

- ANO > variance a kovariance se mění v čase v důsledku změn genetických a prostředí (tj. selekce,…)

3. Neparametrické metody

- obtížně měřitelné znaky
- neznáme fenotyp, známe pořadí
- korelační koeficient dle Spearmana
- stanovíme pořadí rodičů a nezávisle pořadí potomků;
- diference mezi pořadím d_i

$$r_s = 1 - \frac{6.\sum d_i^2}{n.(n^2 - 1)}$$

n - počet dvojic

Př. Použití pořadového korelačního koeficientu podle Spearmana u matek a dcer

matky		dcery	
% tuku	pořadí	% tuku	pořadí
4,6	1	4,4	3
4,5	2	4,0	7
4,4	3	3,6	11
4,3	4	3,9	8
4,2	5	4,6	1
4,1	6	4,3	4
4,0	7	4,5	2
3,9	8	4,2	5
3,8	9	3,5	12
3,7	10	4,1	6
3,6	11	3,7	10
3,5	12	3,8	9

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 198}{12(12^2-1)} = 0{,}3077$$

$$h^2 = R^2 = 0{,}0947$$

Př. Výpočet odhadu koeficientu dědivosti na základě zjištění průměrného pořadí matka – dcera

- Vhodné využití u vlastností, které se nedají přesně číselně vyjádřit nebo při sledování málo početného souboru.
- Užitkovost matek se seřadí podle pořadí od nejvyšší hodnoty užitkovosti do nejnižší a podobně se provede určení pořadí u jejich dcer. Na základě stanovení pořadí u matek přiřadíme ke každé matce pořadí její dcery.
- Soubor se rozdělí na polovinu a vypočítáme průměrné pořadí dcer (**r**) lepších a horších matek a průměrné pořadí lepších a horších matek (**R**).

pořadí	lepší matky									horší matky								
matek	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
dcer	4	3	1	9	2	16	18	15	5	14	12	6	10	7	8	11	13	17

Výpočet odhadu koeficientu dědivosti podle:
- průměrného pořadí dcer a matek:

$$h^2 = 2\frac{\bar{r}_- - \bar{r}_+}{\bar{R}_- - \bar{R}_+} = 0{,}61728$$

- průměrného pořadí dcer:

$$h^2 = 2\frac{2(\bar{r}_- - \bar{r}_+)}{n} = 0{,}61728$$

4. Selekční experiment

Realizovaná dědivost

$$h^2 = \frac{\bar{x}_0 - \bar{x}}{\bar{x}_s - \bar{x}}$$

Realizovaná dědivost v genetickém zisku

$$h^2 = \frac{\Delta G}{d}$$