# Föreläsning 14: Variansanalys

Matematisk statistik

David Bolin
Chalmers University of Technology
Oktober 17, 2018




## Variansanalys

- Vi har med linjär regression visat hur vi kan skatta polynomiella (tex linjära eller kvadratiska) samband mellan en responsvariabel och en eller flera förklarande variabler.
- Vi ska nu undersöka mer generella situationer:
  - Sambandet behöver inte beskrivas av någon enkel funktion
  - Förklarande variabler kan vara både numeriska och kategoriska
- Exempel 1: Vi kan vilja undersöka om personer som har olika bilmärken tenderar att köra olika fort. Vi mäter hastigheten hos $100$ bilar vid en fartkamera och har som förklarande variabel bilmärke.
- Exempel 2: I projektet kan vi vilja undersöka om fiskens vikt påverkar mängden kvicksilver, utan att göra något antagande om linjärt beroende.





## Ensidig variansanalys

- Vi vill studera om en faktor A påverkar en responsvariabel.
- Vi gör totalt $N = \sum_{i=1}^{a} n_i$ mätningar vid $k$ olika faktornivåer:

| Nivå | $1$ | $2$ | $\cdots$ | $k$ |
|---|---|---|---|---|
| | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{k1}$ |
| | $y_{12}$ | $y_{12}$ | $\cdots$ | $y_{k2}$ |
| | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| | $y_{1n_1}$ | $y_{2n_2}$ | $\cdots$ | $y_{kn_a}$ |
| Medel: | $\bar{y}_1$ | $\bar{y}_2$ | $\cdots$ | $\bar{y}_k$ |

- Modell: $Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$
- Vill testa:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1 : \mu_i \neq \mu_j \text{ för några } i, j$$





## Variansanalystabell

| Variation | Kvadratsumma | Frihetsgrader | Medelkvadratsumma | Teststorhet |
|---|---|---|---|---|
| Faktor A | $SS_{Tr} = \sum_{ij}(\bar{y}_i - \bar{y})^2$ | $f_{Tr} = k - 1$ | $MS_{Tr} = SS_{Tr}/f_{Tr}$ | $MS_{Tr}/MS_E$ |
| Residual | $SS_E = \sum_{ij}(y_{ij} - \bar{y}_i)^2$ | $f_E = \sum_i n_i - k$ | $MS_E = SS_E/f_E$ | |
| Total | $SS_{Tot} = SS_E + SS_{Tr}$ | $f = \sum_i n_i - 1$ | | |

- Testet görs med hjälp av ANOVA-tabellen ovan.
- ANOVA står för Analysis of variance.
- $F = MS_A/MS_E$ är $F(k-1, N-k)$-fördelad om $H_0$ är sann. Vi gör alltså hypotestestet med ett vanligt F-test.

## Exempel: Kvicksilver i fisk

| Variation | Kvadratsumma | Frihetsgrader | Medelkvadratsumma | Teststorhet |
|---|---|---|---|---|
| Faktor | 33.3457 | 25 | 1.33 | 124.31 |
| Residual | 2.4679 | 230 | 0.0107 | |
| Total | 35.8137 | 255 | | |

- I projektet tittade vi på hur vikt och provdatum påverkade kvicksilvermätningarna, men vi tittade inte på effekten av var mätningen var gjord.
- ANOVA-tabellen visar resultat från datan från Grupp 1, där jag har valt Station som förklarande variabel och har HG i abborre som responsvariabel.
- Vi jämför teststorheten med en kvantil i F(25,230)-fördelningen: $F_{0.05}(25, 230) = 1.55$. Vi kan förkasta nollhypotesen att infångstplats inte spelar roll!





## Koppling till t-test

Antag att vi endast har två faktornivåer.

- Vi testar då alltså med ANOVA om två populationer (som antas ha samma varians) har samma väntevärde.
- Detta var precis vad vi gjorde med t-test och poolad varians.
- Det visar sig att dessa två tester faktiskt är helt ekvivalenta:
  - När vi gör t-test beräknade vi teststorheten
$$T_{obs} = \frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{1/n_1 + 1/n_2}}$$
  - Vi förkastade $H_0$ om $2(1 - F_{t(n_1+n_2-2)}(|T_{obs}|)) < \alpha$
  - Om vi jämför uttrycket för $s_p^2$ och $MS_E$ i ANOVA-tabellen ser vi $s_p^2 = MS_E$.
  - Med anova förkastar vi $H_0$ om $1 - F_{F(1,n_1+n_2-2)}(F_{obs}) < \alpha$.
  - Om undersöker teststorheten $F_{obs}$ i ANOVA-tabellen ser vi att $F_{obs} = T_{obs}^2$.
  - Om $X \sim t(n)$ så är $X^2 = (Z/\sqrt{V})^2 = Z^2/V \sim F(1, n)$.
  - Alltså kommer p-värdet för F-testet vara detsamma som p-värdet i ett t-test med poolad varians.





## Tabell för tvåsidig variansanalys

| Variation | Kvadratsumma | Frihetsgrader | Medelkvadrat | Teststorhet |
|---|---|---|---|---|
| Faktor A | $SS_A$ | $f_A = a - 1$ | $MS_A = SS_A/f_A$ | $MS_A/MS_E$ |
| Faktor B | $SS_B$ | $f_B = b - 1$ | $MS_B = SS_B/f_B$ | $MS_B/MS_E$ |
| Faktor AB | $SS_{AB}$ | $f_{AB} = f_A f_B$ | $MS_{AB} = SS_{AB}/f_{AB}$ | $MS_{AB}/MS_E$ |
| Residual | $SS_E$ | $f_E = ab(n-1)$ | $MS_E = SS_E/f_E$ | |
| Total | $SS_{Tot}$ | $abn - 1$ | | |

För att utföra de tre hypotesttesterna använder vi att:

1. Om ingen samvariation: $MS_{AB}/MS_E \sim F(f_{AB}, f_E)$.
2. Om ingen effekt av faktor A: $MS_A/MS_E \sim F(f_A, f_E)$.
3. Om ingen effekt av faktor B: $MS_B/MS_E \sim F(f_B, f_E)$.

Utför respektive hypotestest genom att jämföra respektive teststorhet med kritiskt värde i F-fördelingen.