Faculty of Life Sciences

Sammenligning af grupper

Ensidet ANOVA

Claus Ekstrøm
E-mail: ekstrom@life.ku.dk

Program

- Sammenligning af to grupper: tre eksempler
- Sammenligning af mere end to grupper: ensidet ANOVA
  - Data: antibiotika og nedbrydning af organisk materiale
  - Statistisk model
  - Estimation og konfidensintervaller
  - Sammenligning af grupperne (test)
  - Parvise sammenligninger

Case 3, del I: A-vitamin i leveren

A-vitamin tilført på to måder:

- i majsolie (`corn`): $x_1, \ldots, y_{10}$
- i amerikansk olie (`am`): $y_1, \ldots, y_{10}$

Spørgsmål: er A-vitaminkonc. i leveren den samme uanset olietypen?

Statistisk model: alle x'er og y'er er uafhængige og der er ens spredning i de to grupper (samme σ):

$$x_1, \ldots, x_{10} \sim N(\mu_x, \sigma^2), \qquad y_1, \ldots, y_{10} \sim N(\mu_y, \sigma^2)$$

Hypotesen $H_0 : \mu_x = \mu_y$ testes med

$$T = \frac{\hat{\mu}_x - \hat{\mu}_y}{\mathrm{SE}(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\bar{x} - \bar{y}}{\mathrm{SE}(\bar{x} - \bar{y})} = \frac{\bar{x} - \bar{y}}{s\sqrt{1/10 + 1/10}} \sim t_{20-2}$$

R: `t.test(x,y, var.equal=T)`

Case 3, del II: Fiskesmag i lammekød

11 lam i to grupper: 5 lam fik standardfoder ($x_1, \ldots, x_5$), og 6 lam fik standardfoder tilsat fisk ($y_1, \ldots, y_6$).

Spørgsmål: Er der afsmag af fisk i lammekødet?

Statistisk model: alle x'er og y'er er uafhængige, men der er er forskellig spredning i de to grupper:

$$x_1, \ldots, x_5 \sim N(\mu_x, \sigma_x^2), \qquad y_1, \ldots, y_6 \sim N(\mu_y, \sigma_y^2)$$

Hypotesen $H_0 : \mu_x = \mu_y$ testes med

$$T = \frac{\hat{\mu}_x - \hat{\mu}_y}{\mathrm{SE}(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\bar{x} - \bar{y}}{\mathrm{SE}(\bar{x} - \bar{y})} = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/5 + s_y^2/6}} \overset{\text{approx.}}{\sim} t_{\mathrm{df}}$$

hvor frihedsgraderne beregnes udfra s_x og s_y: Se afsnit 5.4, p. 127!

R: `t.test(x,y, var.equal=F)` eller bare `t.test(x,y)`

Opgave 6.3: Fertilitet af lucerne

To klaser fra hver af 10 lucerneplanter: én klase bøjet ned $(x_1, \ldots, x_{10})$, den anden klase eksponeret for sol og vind $(y_1, \ldots, y_{10})$

Her er x'erne og y'erne ikke uafhængige — de kommer parvis fra de samme planter! Vi taler om parvise observationer.

Ser i stedet på differenserne, $d_i = x_i - y_i$.

Statistisk model: d'erne er uafhængige og $d_i \sim N(\mu, \sigma^2)$.

Hypotesen $H_0 : \mu = 0$ testes med et parret t-test:

$$T = \frac{\hat{\mu}}{\mathrm{SE}(\hat{\mu})} = \frac{\bar{d}}{\mathrm{SE}(\bar{d})} = \frac{\bar{d}}{s_d/\sqrt{10}} \sim t_{10-1}$$

R: `t.test(x,y, paired=T)` eller `t.test(x-y)`

Sammenligning af to stikprøver: oversigt

	x, y uafh.?	Samme sd.?	R
A-vitamin	Ja	Ja	`t.test(x,y, var.equal=T)`
Fiskesmag	Ja	Nej	`t.test(x,y)`
Lucerne	Nej		`t.test(x,y, paired=T)`

Når vi skal sammenligne to stikprøver kan vi altså klare os med t-test i forskellige afskyninger.

Hvad hvis vi vil sammenligne tre eller flere stikprøver samtidig? Ensidet ANOVA!

Antibiotika og nedbrydning af organisk materiale

Data

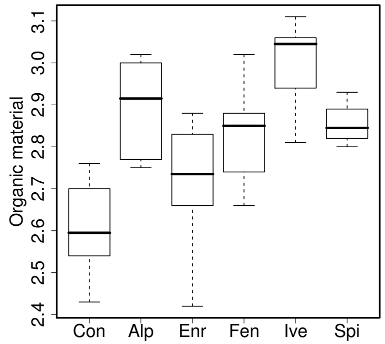
- Fem typer antibiotika og en kontrolbehandling
- 36 kvier inddelt i seks grupper. Foder tilsat antibiotikum
- Gødning gravet ned i poser og mængden af organisk materiale målt efter 8 uger
- For spiramycin: kun fire brugbare målinger

Formål

- Påvirker antibiotika nedbrydningen af organisk materiale?
- Hvis kontrolmålingerne ligger lavere end de andre, tyder det på at antibiotika hæmmer nedbrydningen.
- Ligger de signifikant lavere — eller skyldes det bare tilfældigheder?

Gruppegennemsnit og -spredninger

Type	n_j	$\bar{y}_j$	s_j
Control	6	2.603	0.119
α-cyperm.	6	2.895	0.117
Enrofloxacin	6	2.710	0.162
Fenbendaz.	6	2.833	0.124
Ivermectin	6	3.002	0.109
Spiramycin	4	2.855	0.054



Sammenvejet (pooled) spredningsestimat:

$$s = \sqrt{\frac{1}{28}\left(5 \cdot s_1^2 + \cdots + 3 \cdot s_6^2\right)} = \sqrt{\frac{1}{34-6}\sum_{i=1}^{n}(y_i - \bar{y}_{g(i)})^2} = 0.1217$$

Statistisk model

Husk at $g(i)$ angiver gruppen for observation i. For eksempel

$$g(1) = \cdots = g(6) = \text{control}, \quad g(31) = \cdots = g(34) = \text{Spiramycin}$$
$$g(1) = \cdots = g(6) = 1, \quad g(31) = \cdots = g(34) = 6.$$

Statistisk model: $y_1, \ldots, y_{34}$ er uafhængige og

$$y_i \sim N(\alpha_{g(i)}, \sigma^2)$$

Parametre: $\alpha_1, \ldots, \alpha_6$ og σ.

Ækvivalent formulering:

$$y_i = \alpha_{g(i)} + e_i, \quad e_1, \ldots, e_{34} \sim N(0, \sigma^2) \text{ uafhængige}$$

Statistisk model

Altså:

$$y_i = \alpha_{g(i)} + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

Antagelserne er:

- Alle y_i er normalfordelte
- Middelværdien af y_i er $\alpha_{g(i)}$ — en middelværdi for hver gruppe
- Alle y_i har samme spredning
- Uafhængighed

Estimation og konfidensintervaller

Statistisk model:

$$y_i = \alpha_{g(i)} + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ uafhængige}$$

Parametre: $\alpha_1, \ldots, \alpha_k$ og σ. Især interesseret i forskelle, $\alpha_j - \alpha_l$!

Estimater og estimerede spredninger:

$$\hat{\alpha}_j = \bar{y}_j; \quad \mathrm{SE}(\hat{\alpha}_j) = s\sqrt{1/n_j} = s/\sqrt{n_j}$$
$$\hat{\alpha}_j - \hat{\alpha}_l = \bar{y}_j - \bar{y}_l; \quad \mathrm{SE}(\hat{\alpha}_j - \hat{\alpha}_l) = s\sqrt{1/n_j + 1/n_l}$$
$$\hat{\sigma} = s$$

Konfidensintervaller på sædvanlig vis:

$$\text{estimat} \pm t_{0.975, n-k} \cdot \mathrm{SE}(\text{estimat})$$

NB. s bruges også ved sammenligning af to af grupperne!

Ensidet ANOVA i R

Fit af ensidet ANOVA model:

```
> model1 <- lm(org~factor(type))
> summary(model1)
```

R vælger en referencegruppe — den første efter alfabetisk rækkefølge — og estimerer forskelle i forhold til denne gruppe.

Vi vil hellere bruge kontrolgruppen som reference:

```
> type <- relevel(type, ref="Control")
> model1 <- lm(org~factor(type))
> summary(model1)
```

Ensidet ANOVA i R

Output fra `summary(model1)`:

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.60333    0.04970  52.379  < 2e-16 ***
factor(type)Alfacyp   0.29167    0.07029   4.150 0.000281 ***
factor(type)Enroflox  0.10667    0.07029   1.518 0.140338
factor(type)Fenbenda  0.23000    0.07029   3.272 0.002834 **
factor(type)Ivermect  0.39833    0.07029   5.667  4.5e-06 ***
factor(type)Spiramyc  0.25167    0.07858   3.202 0.003384 **

Residual standard error: 0.1217 on 28 degrees of freedom
```

Fortolkninger:

- Estimat og CI for α_{cont}, $\alpha_{\text{Fenb}} - \alpha_{\text{cont}}$ og α_{Fenb}? Estimat for σ?
- Hvorfor er der forskellige SE'er?

Ensidet ANOVA i R

Hvis vi hellere vil have gruppegennemsnit isf. forskelle til kontrolgruppen:

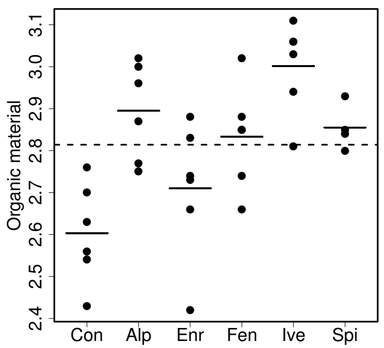
```
> model2 <- lm(org~factor(type)-1)
> summary(model2)
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
factor(type)Control   2.60333    0.04970   52.38   <2e-16 ***
factor(type)Alfacyp   2.89500    0.04970   58.25   <2e-16 ***
factor(type)Enroflox  2.71000    0.04970   54.53   <2e-16 ***
factor(type)Fenbenda  2.83333    0.04970   57.01   <2e-16 ***
factor(type)Ivermect  3.00167    0.04970   60.39   <2e-16 ***
factor(type)Spiramyc  2.85500    0.06087   46.90   <2e-16 ***

Residual standard error: 0.1217 on 28 degrees of freedom
```

De to specifikationer er gode til hver sit formål!

Hypotese. Variation indenfor og mellem grupper

Hypotese, $H_0 : \alpha_1 = \cdots = \alpha_k$.

Alternativ, H_A : mindst to α'er er forskellige.



- Variation indenfor grupper — punkter vs. fuldt optrukne liniestykker
  $$\text{SS}_e = \sum_{i=1}^{n} (y_i - \bar{y}_{g(i)})^2$$
- Variation mellem grupper — Fuldt optrukne linieestykker vs. stiplet linie
  $$\text{SS}_{\text{grp}} = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$
- Teststørrelse
  $$F = \frac{\text{MS}_{\text{grp}}}{\text{MS}_e} = \frac{\text{SS}_{\text{grp}}/(k-1)}{\text{SS}_e/(n-k)}$$

Sammenligning af alle grupperne

Kan kun bruge `model1` til dette — *ikke* `model2` med `-1`!

```
> anova(model1)
             Df  Sum Sq Mean Sq F value    Pr(>F)
factor(type)  5 0.59082 0.11816  7.9726 8.953e-05 ***
Residuals    28 0.41500 0.01482
```

Teststørrelse

$$F = \frac{\text{MS}_{\text{grp}}}{\text{MS}_e} = \frac{\text{SS}_{\text{grp}}/(k-1)}{\text{SS}_e/(n-k)}$$

Store værdier af F er kritiske — passer dårligt med hypotesen, så

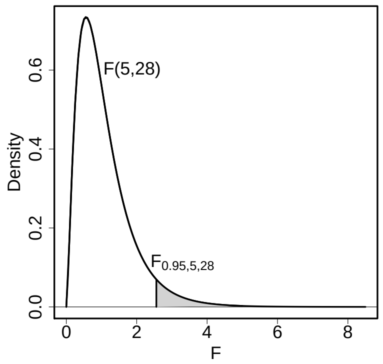
$$p = P(F \geq F_{\text{obs}}) = P(F \geq 7.97) = 0.00009$$

så der er med stor sikkerhed påvist en forskel på typerne.

Hvordan kom vi frem til p-værdien?

F-fordelingen

Hvis hypotesen er sand er F-teststørrelsen F-fordelt med $(k-1, n-k)$ frihedsgrader.



$p = P(F \geq 7.97) = 0.00009$

Vi afviser H_0 hvis F_{obs} er større end 95%-fraktilen, her $F_{0.95,5,28} = 2.56$.

Sandsynligheder og fraktiler i R:

```
> pf(7.97, df1=5, df2=28)
[1] 0.9999102
> qf(0.95, df1=5, df2=28)
[1] 2.558128
```


Parvise sammenligninger

Antag at vi er specielt interesseret i forskel mellem kontrolgruppen (gruppe 1) og Fenbendazolegruppen (gruppe 4): $\alpha_4 - \alpha_1$.

Estimat og estimeret spredning:

$$\hat{\alpha}_4 - \hat{\alpha}_1 = 2.833; \quad \mathrm{SE}(\hat{\alpha}_4 - \hat{\alpha}_1) = 0.07029$$

- Konfidensinterval for $\alpha_4 - \alpha_1$?
- Test for hypotesen $H_0 : \alpha_1 = \alpha_4$?
- Er alle grupperne signifikant forskellige fra kontrolgruppen?

LSD-værdi: least significant difference

Hvor stort skal estimatet for forskellen mellem to grupper være for at den bliver signifikant?

Forskellen $\hat{\alpha}_j - \hat{\alpha}_l$ er signifikant hvis og kun hvis

$$|T| = \frac{|\hat{\alpha}_j - \hat{\alpha}_l|}{\mathrm{SE}(\hat{\alpha}_j - \hat{\alpha}_l)} > t_{0.975,n-k} \Leftrightarrow |\hat{\alpha}_j - \hat{\alpha}_l| > t_{0.975,n-k} \cdot \mathrm{SE}(\hat{\alpha}_j - \hat{\alpha}_l)$$

Altså er den mindste signifikante forskel:

$$\mathrm{LSD}_{j,l} = t_{0.975,n-k} \cdot \mathrm{SE}(\hat{\alpha}_j - \hat{\alpha}_l) = t_{0.975,n-k} \cdot s \cdot \sqrt{1/n_j + 1/n_l}$$

LSD for kontrol og fenbend.: $2.048 \cdot 0.1217 \cdot \sqrt{1/6 + 1/6} = 0.144$

Hvis n' obs. i alle grupper: samme LSD-værdi for alle par af grupper:

$$\mathrm{LSD} = t_{0.975,n-k} \cdot \mathrm{SE}(\hat{\alpha}_j - \hat{\alpha}_l) = t_{0.975,n-k} \cdot s \cdot \sqrt{2/n'}$$

Konklusion

Vi har med stor sikkerhed påvist at der er forskel på antibiotikatyperne ($p < 0.0001$)

For alle typer på nær Enrofloxacin er mængden af organisk materiale signifikant højere end for kontrolgruppen.

Angiv desuden estimater og konfidensintervaller for α'er og/eller for forskelle til kontrolgruppen.

Resumé: ensidet variansanalyse

- Statistisk model: normalfordeling, ens spredning i gruppernem uafhængighed
- Estimation: gruppegennemsnit og sammenvejet stikprøvespredning
- Konfidensinterval: $\mathrm{estimat} \pm t_{0.975,n-k} \cdot \mathrm{SE}(\mathrm{estimat})$
- Hypotesen om ens middelværdier testes med $F = \mathrm{MS}_{\mathrm{grp}}/\mathrm{MS}_e$.
- Parvise sammenligninger foretages "indenfor" modellen, således at alle observationer bruges til at estimere spredningen.

Hvis der kun er to grupper, så kan vi klare os med t-test.
Forskellige "versioner":

- Er stikprøverne uafhængige?
- Kan spredningerne antages at være ens?

Dagens hovedpunkter

- Ensidet variansanalyse
- Antagelser for ensidet variansanalyse
- Hypoteser for ensidet variansanalyse
- Teststørrelse og F-fordelingen

På onsdag:

- Modelkontrol og prædiktion
- Sammenhængen mellem modellere: ligheder og forskelle
- Eksempler og hængepartier
- Uge 5: Multipel regression og tosidet ANOVA.

Ordliste

Engelsk	Dansk
LSD	Mindste signifikante forskel (LSD)
one-way ANOVA	ensidet variansanalyse
pooled	sammenvejet
variation between groups	variation mellem grupper
variation within groups	variation indenfor grupper