Faculty of Life Sciences

Ensidet variansanalyse
Normalfordelingen

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

Program

- Ensidet variansanalyse (one-way ANOVA)
  - Hvilken type data? Hvad er problemstillingen?
  - Variation mellem grupper og indenfor grupper
  - Residualer
- Normalfordelingen
  - Histogram og tæthed
  - Sandsynligheder
  - Symmetri, centrum og spredning

Antibiotika og nedbrydning af organisk materiale

Data

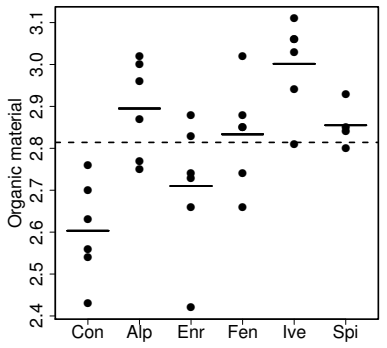
- Fem typer antibiotika og en kontrolbehandling
- 36 kvier inddelt i seks grupper. Foder tilsat antibiotikum
- Gødning gravet ned i poser og mængden af organisk materiale målt efter 8 uger
- For spiramycin: kun fire brugbare målinger

Formål

- Påvirker antibiotika nedbrydningen af organisk materiale?
- Hvis kontrolmålingerne ligger lavere end de andre, tyder det på at antibiotika hæmmer nedbrydningen.
- Men hvor meget lavere skal de ligge for at vi kan drage den konklusion? Det får vi ikke svar på i dag...

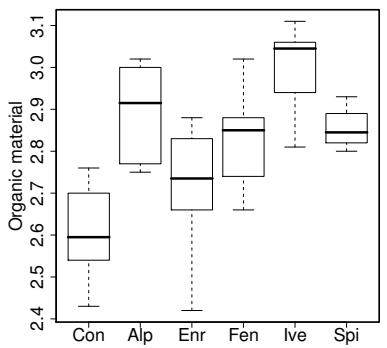
Tegninger

Data



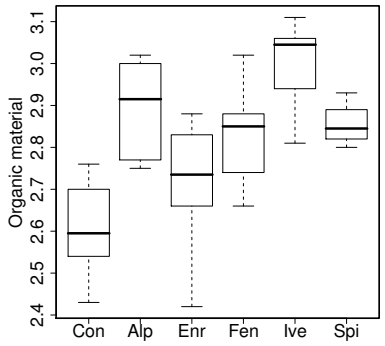
Parallelle boxplot

```
boxplot(org~treat)
```



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



- Hvad kan vi se fra tegninger og tal?
- Kan vi konkludere at der er forskel på grupperne?

Populationer, stikprøver og estimater

Population vs. stikprøve

- De 34 kvier er en stikprøve fra populationen af kvier
- Faktisk forestiller vi os seks delpopulationer: kvier der får behandling 1, kvier der får behandling 2, osv.
- En kvie fra gruppe j er repræsentativ for den pågældende population
- Vil drage konklusioner om populationerne på grundlag af stikprøverne

Middelværdi/gennemsnit i population:

- α_j er populationsgennemsnit for kvier fra gruppe j
- Stikprøvegennemsnittet $\bar{y}_j$ er estimat for α_j: $\hat{\alpha}_j = \bar{y}_j$
- Hvordan kan vi udtrykke at der ikke er nogen effekt af antibiotika?

Notation

- k er antal grupper, her $k = 6$
- n_j er antal obs. i gruppe j, her $n_1 = \cdots = n_5 = 6$, $n_6 = 4$.
- $g(i)$ angiver gruppen for observation i. For eksempel

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

eller

$$g(1) = \cdots = g(6) = 1, \qquad g(31) = \cdots = g(34) = 6.$$

- Stikprøvegennemsnit og -spredning i gruppe j:

$$\bar{y}_j = \frac{1}{n_j} \sum_{i:g(i)=j} y_i \qquad s_j = \sqrt{\frac{1}{n_j - 1} \sum_{i:g(i)=j} (y_i - \bar{y}_j)^2}$$

Altså: $\bar{y}_j$ er gennemsnit af de observationer i der har $g(i) = j$, dvs. kommer fra gruppe j.

Sammenvejet stikprøvespredning

Hvis der er nogenlunde samme variation i grupperne:

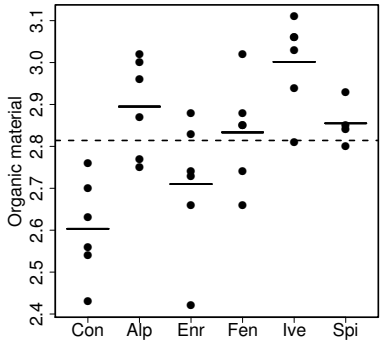
- beregning af et fælles stikprøvespredning
- ... som er et estimat for en fælles spredning i populationerne

Sammenvejet — eller pooled — stikprøvespredning:

$$\begin{aligned} s &= \sqrt{\frac{1}{n-k} \sum_{j=1}^{k} (n_j - 1) s_j^2} \\ &= \sqrt{\frac{1}{28} \left(5 \cdot s_1^2 + 5 \cdot s_2^2 + \cdots + 3 \cdot s_6^2\right)} = 0.1217 \end{aligned}$$

Bemærk: Varianserne lægges sammen — ikke spredningerne.

Variation indenfor og mellem grupper



- Variation indenfor grupper — punkter vs. fuldt optrukne liniestykker

$$SS_e = \sum_{i=1}^{n} \left(y_i - \bar{y}_{g(i)}\right)^2$$

- Variation mellem grupper — Fuldt optrukne linieestykker vs. stiplet linie

$$SS_{grp} = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

- Total variation — Punkter vs. stiplet linie

$$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Variationen mellem grupper skal ses ift. variationen indenfor grupper!

Residualer mm.

Husk residualer fra lineære regression: $r_i = y_i - \hat{\alpha} - \hat{\beta} \cdot x_i$.

Ensidet ANOVA:

- Residualer (hvor meget "skyder vi galt"?)

$$r_i = y_i - \bar{y}_{g(i)} = \text{observation} - \text{estimat}$$

- Residualkvadratsummen er netop SS_e:

$$SS_e = \sum_{i=1}^{n} \left(y_i - \bar{y}_{g(i)}\right)^2 = \sum_{i=1}^{n} r_i^2$$

- Beregning af spredningsestimat udfra residualkvadratsum:

$$s = \sqrt{\frac{1}{n-k} \sum_{i=1}^{n} r_i^2} = \sqrt{\frac{1}{df_e} \sum_{i=1}^{n} r_i^2}$$

Sådan er det altid!

Uparrede og parrede stikprøver

Uparrede forsøgsdesigns: 2 grupper — ensidet variansanalyse.

Parrede forsøgsdesign: Hvad gør vi her? Hvad vil vi? Hvorfor er parrede forsøg "smarte"?

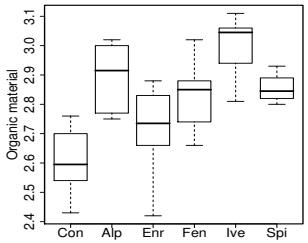
Ensidet variansanalyse: resumé

- Observationerne inddelt i k grupper, fx. svarende til forskellige behandlinger, sorter, aldersgrupper,...
- Formål: sammenligning af grupperne
- Opdeling af total variation i variation mellem grupper og variation indenfor grupper
- Sammenvejet spredningsestimat, s
- Kan ikke konkludere om der er forskel på grupperne udfra tegninger og gruppegennemsnit alene. Vi skal have en statistisk model for data!

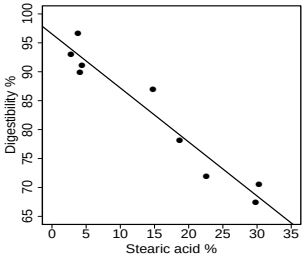
Hvorfor skal vi lære om normalfordelingen (nu)?

Har set tre typer af data/eksperimenter med kontinuerte data:

Ensidet ANOVA



Lineær regression



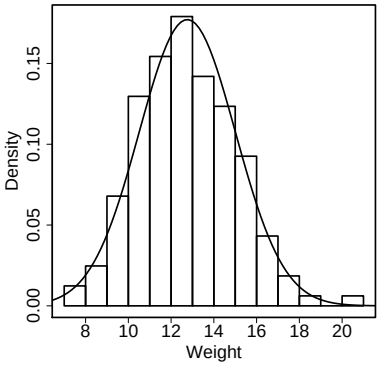
En stikprøve:

Blood pressure								
96	119	119	108	126	128	110	105	94

Vi skal bruge normalfordelingen for alle tre forsøgstyper/datatyper!

Vægt af krabber

- 162 krabber på en bestemt alder vejet: $y_1, \ldots, y_{162}$.
- R: $\bar{y} = 12.76$, $s = 2.25$
- Histogram normeret så det samlede areal af rektangler er 1
- Graf for f, hvor f er tætheden for normalfordelingen



$$f(y) = \frac{1}{\sqrt{2\pi \cdot 2.25^2}} \exp\left(-\frac{1}{2 \cdot 2.25^2}(y - 12.76)^2\right)$$

Grafen for f er en fin approksimation af histogrammet.

Sandsynligheder

Husk: for standardiseret histogram er "relativ frekvens = areal af rektangel", fx.

$$\frac{\text{antal krabber mellem 14 g og 15 g}}{162} = 0.12$$

Tilsvarende for tætheden: sandsynligheden for at en observation falder i intervallet fra a til b er lig arealet under kurven, fx.

$$P(14 < Y < 15) = \int_a^b f(y)\,dy = 0.13$$

De to sandsynligheder er ikke ens: population vs. stikprøve.

- Hvis populationsværdier er fordelt som tætheden beskriver, så *vil* histogram for stikprøve fra populationen ligne tætheden
- Normalfordelingstæthed "som model for" histogrammet

Normalfordeling med middelværdi μ og spredning σ

Udskift tallene 12.76 og 2.25 med μ og $\sigma > 0$:

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

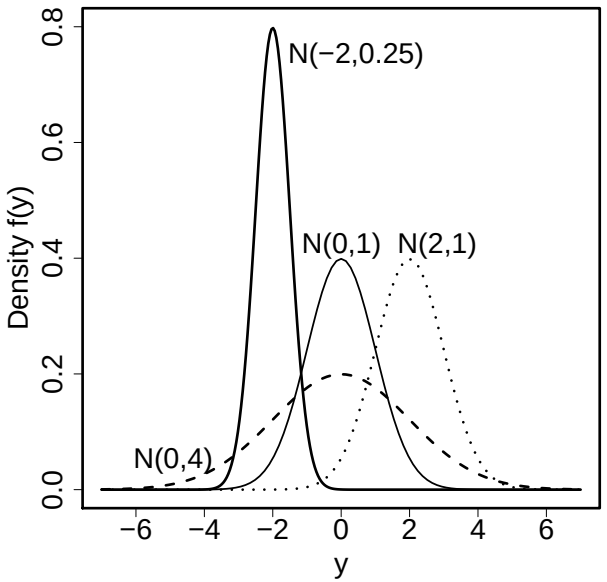
Vi siger at en variabel Y er normalfordelt med middelværdi μ og spredning σ hvis

$$P(a < Y < b) = \int_a^b f(y)\,dy.$$

for alle a og b, dvs. for alle intervaller. Vi skriver $Y \sim N(\mu, \sigma^2)$.

Bemærk: σ^2 — ikke σ i denne notation. Altså: hvis Y er normalfordelt med middelværdi 3 og spredning 2, så er $Y \sim N(3, 4)$.

Symmetri — centrum — spredning



Dagens hovedpunkter

- Ensidet variansanalyse: sammenligning af middelværdi for k grupper
- Variation mellem grupper / variation indenfor grupper
- Population og tæthed vs. stikprøve og histogram
- Sandsynlighed lig areal under tæthed
- Tæthed for normalfordeling: symmetri, centrum og spredning

På onsdag (og måske næste mandag):

- Egenskaber for normalfordelingen
- Hvordan kontrollerer vi, at data er normalfordelt?
- Hvorfor lige netop normalfordelingen?

Tyvstarter ved øvelserne i dag med at undersøge egenskaber vha. simulation af N-fordelte variable (opgave 4.3).