# CHAPTER 6. ANALYSIS OF VARIANCE (ONE-WAY)

6.1. What is “Analysis of variance” (ANOVA)? 6.2.

One-Way Analysis of Variance

6.3. The Least significant difference intervals

6.4. The p-value

# Statistisk Dataanalyse

## Chapter 6: Assignments

1. We want to investigate if there is an effect of the type of fertilizer applied to apple trees and the production of apples. We randomly select 15 trees and randomly assign them to one of three groups (5 trees per group). We perform a test in which we apply one type of fertilizer (fertilizer 1, 2 or 3) to each group. The data are shown. At a α=0.05, can it be concluded that there is a significant difference in the production of apples depending on which fertilizer is used? Which fertilizer/fertilizers causes a higher/lower production than the other/others?

| Fertilizer 1 | Fertilizer 2 | Fertilizer 3 |
|---|---|---|
| 10 | 6 | 5 |
| 12 | 8 | 9 |
| 9 | 3 | 12 |
| 15 | 0 | 8 |
| 13 | 2 | 4 |

- Vi starter allerførst med at udregne Middelværdi for hver kolonne af Kunstgødning.

| | |
|---|---|
| Summen á Fertilizer 1 | $10 + 12 + 9 + 15 + 13 = 59$ |
| Middelværdi | $\frac{59}{5} = 11,8$ |

| | |
|---|---|
| Summen á Fertilizer 2 | $6 + 8 + 3 + 0 + 2 = 19$ |
| Middelværdi | $\frac{19}{5} = 3,8$ |

| | |
|---|---|
| Summen á Fertilizer 3 | $5 + 9 + 12 + 8 + 4 = 38$ |
| Middelværdi | $\frac{38}{5} = 7,6$ |

- Nu skal vi til at udregne selve den totale middelværdi, som er vores "sum of mean" og det betyder at vi lægger alle middelværdierne sammen til en.

| | |
|---|---|
| Summen af Kunstgødernes Middelværdi | $11,8 + 3,8 + 7,6 \approx 23,2$ |
| Totale Middelværdi af Kunstgøderne | $\frac{23,2}{3} \approx 7,733333$ |

- Nu skal vi udregne summen af kvadraterne for alle kunstgøderne.
- Dette betyder, at vi tager den enkelte dataværdi og trækker det fra den totale middelværdi af kunstgøderne. Vi viser den enkelte eksempel herhenne.

| | |
|---|---|
| Sum of Square Total | $(10-7{,}73)^2+(12-7{,}73)^2+(9-7{,}73)^2+(15-7{,}73)^2+(13-7{,}73)^2+(6-7{,}73)^2+(8-7{,}73)^2+(3-7{,}73)^2+(0-7{,}73)^2+(2-7{,}73)^2+(5-7{,}73)^2+(9-7{,}73)^2+(12-7{,}73)^2+(8-7{,}73)^2+(4-7{,}73)^2=264{,}9335$ |
| Sum of Square Between | $5\cdot(11{,}8-7{,}73)^2=82{,}8245$ |
| Sum of Square Between | $5\cdot(3{,}8-7{,}73)^2\approx 77{,}2245$ |
| Sum of Square Between | $5\cdot(7{,}6-7{,}73)^2\approx 0{,}0845$ |
| Sum of Total Square Between | $82{,}8245+77{,}2245+0{,}0845=160{,}1335$ |

- Nu skal vi udregne Sum Square Residual og dette er muligt ved at trække Summen af Square-Total fra Summen af Sum of Total Square Between

| | |
|---|---|
| SS-Total - SS-Between | $\mathbf{264{,}9335-160{,}1335\approx 104{,}8}$ |

- Nu skal vi udregne de adskillige ting som er vist i tabellen på Præsentationen.

| X | Sum of Squares | d.f. | Mean | F-Ratio |
|---|---|---|---|---|
| Factor | SS-Between = 160,1335 | Antal gruppe - 1 => 3-1 = 2 | Dividere Total Square Between med 2 $\frac{160{,}1335}{2}=80{,}06675$ | Dividere Mean Faktor med Mean Residual $\frac{80{,}06675}{8{,}733333}\approx 9{,}167949$ |
| Residual | $\mathbf{264{,}9335-160{,}1335\approx 104{,}8}$ | $14-2=12$ | Dividere SS-Between Residual med df-Residual $\frac{104{,}8}{12}\approx 8{,}733333$ | |
| Total | SS-Total = 264,9335 | Antal dataværdier => 15-1 = 14 | | |

- Nu, hvor frihedsgraden er blevet fundet, skal vi derfor finde den kritiske værdi som er markeret med blå.

| | |
|---|---|
| k-1, N-k | (3,89) |

- Vi kan se, at selve F-Ratio er større end den Kritiske Værdi og derfor kan vi sige at det er følgende:
- F-Ratio > F-Tabel
- Fordi F-Ratio er Større end F-Tabel, kan vi afvise Null-Hypotesen, hvilket betyder at der er en forskel mellem middelværdier af de 3 Kunstgødninger.
- Derfor skal vi nu anvende LSD-Testen for at kunne finde ud af "specifikt", hvilken Kunstgødning adskiller sig fra den ene og om der er andre som ligner det samme.

| F-Ratio > F-Tabel | | | LSD-Intervaller |
|---|---|---|---|
| Kunstgødningerne | Middelværdi | Nedre Grænse | Øvre Grænse |
| Type 1 | 11,8 | 9,76 | 13,83 |
| Type 2 | 3,8 | 1,76 | 5,83 |
| Type 3 | 7,6 | 5,56 | 9,63 |
| (95%,12) | | | |

- Vi kan her til sidst konkludere, at der er en signifikant forskel mellem de 3 forskellige kunstgødningstyper. Vi kunne eksempelvis gennem LSD-Testen se, at Type 1 var den som adskilte sig mest fra Type 2 og 3, da den producerede flest æbler end Type og 3 i forhold til dets LSD-Interval.

2. We want to evaluate three different methods to lower the blood pressure of individuals that have been diagnosed with high blood pressure. Eighteen subjects are randomly assigned to three groups (6 per group): the first group takes medication, the second group exercises, and the third one follows a specific diet. After four weeks, the reduction in each person's blood pressure is recorded. Is there a significant difference among the reduction obtained from each of the three methods? If yes, which method was more effective?

| Medication | Exercise | Diet |
|---|---|---|
| 12 | 5 | 6 |
| 8 | 9 | 10 |
| 11 | 2 | 5 |
| 17 | 0 | 9 |
| 16 | 1 | 8 |
| 15 | 3 | 6 |

# Statistisk Dataanalyse

- Vi starter allerførst med at udregne Middelværdi for hver kolonne af Forsøg på Reducering af Blodtryk.

| | |
|---|---|
| Summen á Medikament | $12 + 8 + 11 + 17 + 16 + 15 = 79$ |
| Middelværdi | $\frac{79}{6} \approx 13,16667$ |

| | |
|---|---|
| Summen á Træning | $5 + 9 + 2 + 0 + 1 + 3 = 20$ |
| Middelværdi | $\frac{20}{6} = \frac{10}{3} \approx 3,333333$ |

| | |
|---|---|
| Summen á Kost | $6 + 10 + 5 + 9 + 8 + 0 = 38$ |
| Middelværdi | $\frac{38}{6} = \frac{19}{3} \approx 6,333333$ |

- Nu skal vi til at udregne selve den totale middelværdi, som er vores "sum of mean" og det betyder at vi lægger alle middelværdierne sammen til en.

| | |
|---|---|
| Summen af Forsøgernes Middelværdi | $13,16 + 3,33 + 6,33 = 22,82$ |
| Totale Middelværdi af Forsøgerne | $\frac{22,82}{3} \approx 7,606667$ |

- Nu skal vi udregne summen af kvadraterne for alle Forsøg.
- Dette betyder, at vi tager den enkelte dataværdi og trækker det fra den totale middelværdi af kunstgøderne. Vi viser den enkelte eksempel herhenne.

| | |
|---|---|
| Sum of Square Total | $(12 - 7,61)^2 + (8 - 7,61)^2 + (11 - 7,61)^2 + (17 - 7,61)^2 + (16 - 7,61)^2 + (15 - 7,61)^2 + (5 - 7,61)^2 + (9 - 7,61)^2 + (2 - 7,61)^2 + (0 - 7,61)^2 + (1 - 7,61)^2 + (3 - 7,61)^2 + (6 - 7,61)^2 + (10 - 7,61)^2 + (5 - 7,61)^2 + (9 - 7,61)^2 + (8 - 7,61)^2 + (6 - 7,61)^2 = 426,9578$ |
| Sum of Square Between | $6 \cdot (13,16 - 7,61)^2 = 184,815$ |
| Sum of Square Between | $6 \cdot (3,33 - 7,61)^2 \approx 109,9104$ |
| Sum of Square Between | $6 \cdot (7,33 - 7,61)^2 \approx 0,4704$ |
| Sum of Total Square Between | $184,815 + 109,9104 + 0,4704 \approx 295,1958$ |

- Nu skal vi udregne Sum Square Residual og dette er muligt ved at trække Summen af Square-Total fra Summen af Sum of Total Square Between

| | |
|---|---|
| SS-Total - SS-Between | $426,9578 - 295,1958 = 131,762$ |

- Nu skal vi udregne de adskillige ting som er vist i tabellen på Præsentationen.

| X | Sum of Squares | d.f. | Mean | F-Ratio |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **Factor** | SS-Between = 295,1958 | Antal gruppe - 1 => 3-1 = 2 | Dividere Total Square Between med 2 $\frac{295,1958}{2} = 147,5979$ | Dividere Mean Faktor med Mean Residual $\frac{147,5979}{8,235} \approx 17,92324$ |
| **Residual** | $\mathbf{426,9578 - 295,1958 = 131,762}$ | $18 - 2 = 16$ | Dividere SS-Between Residual med df-Residual $\frac{131,76}{16} = 8,235$ | |
| **Total** | SS-Total = 426,9589 | Antal dataværdier => 18-1 = 17 | | |

- Nu, hvor frihedsgraden er blevet fundet, skal vi derfor finde den kritiske værdi som er markeret med blå.

| | |
|---|---|
| k-1, N-k | (3,68) |

- Vi kan se, at selve F-Ratio er større end den Kritiske Værdi og derfor kan vi sige at det er følgende:
- F-Ratio > F-Tabel
- Fordi F-Ratio er Større end F-Tabel, kan vi afvise Null-Hypotesen, hvilket betyder at der er en forskel mellem middelværdier af de 3 Kunstgødninger.
- Derfor skal vi nu anvende LSD-Testen for at kunne finde ud af "specifikt", hvilken Forsøg adskiller sig fra den ene og om der er andre som ligner det samme.

| F-Ratio > F-Tabel | | | LSD-Intervaller |
|---|---|---|---|
| Forsøg | Middelværdi | Nedre Grænse | Øvre Grænse |
| Medikament | 13,16 | 11,00 | 15,31 |
| Træning | 3,33 | 1,17 | 5,48 |
| Kost | 6,33 | 4,17 | 8,48 |
| (95%,12) | | | |

**Vi kan her til sidst konkludere, at der er en signifikant forskel mellem de 3 forsøger til at reducere blodtrykket. I tilfældet, kan det ses at Medicin/Medikamentet er det mest effektive, da det har det højeste middelværdi sammenlignet med de andre metoder. Vi kan se, at der er ingen effektivitet i brug af metoder som Træning og Kost, da de resulterer den samme effektivitet. Derfor kan det siges, at Medikament er noget anderledes.**

3. In the table below, there are randomly selected scores for eight amateur basketball teams in each of five Danish regions, for a particular weekend. Is there sufficient evidence to support that there is a difference in mean scores by region? If yes, which region/s got the highest scores and which one the lowest?

# Statistisk Dataanalyse

| Region Hovedstaden | Region Sjælland | Region Syddanmark | Region Midtjylland | Region Nordjylland |
|---|---|---|---|---|
| 68 | 78 | 89 | 62 | 57 |
| 75 | 79 | 87 | 74 | 65 |
| 95 | 65 | 75 | 71 | 78 |
| 85 | 67 | 65 | 70 | 88 |
| 84 | 60 | 84 | 72 | 67 |
| 88 | 79 | 92 | 72 | 77 |
| 85 | 57 | 84 | 64 | 72 |
| 75 | 74 | 72 | 75 | 69 |

- Vi starter allerførst med at udregne Middelværdi for hver kolonne af de danske regioner.

| | |
|---|---|
| Summen á Hovedstad | $\mathbf{68 + 75 + 95 + 85 + 84 + 88 + 85 + 75 = 655}$ |
| Middelværdi | $\frac{\mathbf{655}}{\mathbf{8}} = \mathbf{81,875}$ |

| | |
|---|---|
| Summen á Sjælland | $\mathbf{78 + 79 + 65 + 67 + 60 + 79 + 57 + 74 = 559}$ |
| Middelværdi | $\frac{\mathbf{559}}{\mathbf{8}} = \mathbf{69,875}$ |

| | |
|---|---|
| Summen á Syddanmark | $\mathbf{89 + 87 + 75 + 65 + 84 + 92 + 84 + 72 = 648}$ |
| Middelværdi | $\frac{\mathbf{648}}{\mathbf{8}} = \mathbf{81}$ |

| | |
|---|---|
| Summen á Midtjylland | $\mathbf{62 + 74 + 71 + 70 + 72 + 72 + 64 + 75 = 560}$ |
| Middelværdi | $\frac{\mathbf{560}}{\mathbf{8}} = \mathbf{70}$ |

| | |
|---|---|
| Summen á Nordjylland | $\mathbf{57 + 65 + 78 + 88 + 67 + 77 + 72 + 69 = 573}$ |
| Middelværdi | $\frac{\mathbf{573}}{\mathbf{8}} = \mathbf{71,625}$ |

- Nu skal vi til at udregne selve den totale middelværdi, som er vores "sum of mean" og det betyder at vi lægger alle middelværdierne sammen til en.

| | |
|---|---|
| Summen af Regionernes Middelværdi | $\mathbf{81,875 + 69,875 + 81 + 70 + 71,625 = 374,375}$ |
| Totale Middelværdi af Regionerne | $\frac{\mathbf{374,375}}{\mathbf{5}} = \mathbf{74,875}$ |

- **Nu skal vi udregne summen af kvadraterne for alle Forsøg.**
- Dette betyder, at vi tager den enkelte dataværdi og trækker det fra den totale middelværdi af kunstgøderne. Vi viser den enkelte eksempel herhenne. 74,875

| | |
|---|---|
| Sum of Square Total | $(68-74{,}875)^2+(75-74{,}875)^2+(95-74{,}875)^2+(85-74{,}875)^2+(84-74{,}875)^2+(88-74{,}875)^2+(85-74{,}875)^2+(75-74{,}875)^2+(78-74{,}875)^2+(79-74{,}875)^2+(65-74{,}875)^2+(67-74{,}875)^2+(60-74{,}875)^2+(79-74{,}875)^2+(57-74{,}875)^2+(74-74{,}875)^2+(89-74{,}875)^2+(87-74{,}875)^2+(75-74{,}875)^2+(65-74{,}875)^2+(84-74{,}875)^2+(92-74{,}875)^2+(84-74{,}875)^2+(72-74{,}875)^2+(62-74{,}875)^2+(74-74{,}875)^2+(71-74{,}875)^2+(70-74{,}875)^2+(72-74{,}875)^2+(72-74{,}875)^2+(64-74{,}875)^2+(75-74{,}875)^2+(57-74{,}875)^2+(65-74{,}875)^2+(78-74{,}875)^2+(88-74{,}875)^2+(67-74{,}875)^2+(77-74{,}875)^2+(72-74{,}875)^2+(69-74{,}875)^2 = 3618{,}375$ |
| Sum of Square Between | $8\cdot(81{,}875-74{,}875)^2=392$ |
| Sum of Square Between | $8\cdot(74{,}875-69{,}875)^2=200$ |
| Sum of Square Between | $8\cdot(81-74{,}875)^2=300{,}125$ |

| | |
|---|---|
| Sum of Square Between | $8 \cdot (70 - 74{,}875)^2 = 190{,}125$ |
| Sum of Square Between | $8 \cdot (71{,}625 - 74{,}875)^2 = 84{,}5$ |
| Sum of Total Square Between | $392 + 200 + 300{,}125 + 190{,}125 + 84{,}5 = 1166{,}75$ |

- Nu skal vi udregne Sum Square Residual og dette er muligt ved at trække Summen af Square-Total fra Summen af Sum of Total Square Between

| | |
|---|---|
| SS-Total - SS-Between | $\mathbf{3618{,}375 - 1166{,}75 = 2451{,}625}$ |

- Nu skal vi udregne de adskillige ting som er vist i tabellen på Præsentationen.

| X | Sum of Squares | d.f. | Mean | F-Ratio |
|---|---|---|---|---|
| Factor | SS-Between = 1166,75 | Antal gruppe - 1 => 5-1 = 4 | Dividere Total Square Between med 2 $\frac{1166{,}75}{4} = 291{,}6875$ | Dividere Mean Faktor med Mean Residual $\frac{291{,}68}{70} \approx 4{,}166857$ |
| Residual | $\mathbf{3618{,}375 - 1166{,}75 = 2451{,}625}$ | $40 - 4 = 36$ | Dividere SS-Between Residual med df-Residual $\frac{2451{,}625}{35} \approx 70{,}04643$ | |
| Total | SS-Total = 3618,375 | Antal dataværdier => 40-1 = 39 | | |

- Nu, hvor frihedsgraden er blevet fundet, skal vi derfor finde den kritiske værdi som er markeret med blå.

| | |
|---|---|
| k-1, N-k | (2,63) |

- Vi kan se, at selve F-Ratio er større end den Kritiske Værdi og derfor kan vi sige at det er følgende:
- F-Ratio > F-Tabel
- Fordi F-Ratio er Større end F-Tabel, kan vi afvise Null-Hypotesen, hvilket betyder at der er en forskel mellem middelværdier af de 3 Kunstgødninger.
- Derfor skal vi nu anvende LSD-Testen for at kunne finde ud af "specifikt", hvilken Region adskiller sig fra den ene i forhold til Score.

| F-Ratio > F-Tabel | | | LSD-Intervaller |
|---|---|---|---|
| Score-Områder | Middelværdi | Nedre Grænse | Øvre Grænse |
| Hovedstaden | 81,875 | $81{,}875 - \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = 81{,}875 - \frac{2{,}445605}{2^{\frac{3}{2}}} \approx 81{,}01035$ | $81{,}875 + \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = 81{,}875 + \frac{2{,}445605}{2^{\frac{3}{2}}} \approx 82{,}73965$ |
| Sjælland | 69,875 | $69{,}875 - \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = 69{,}875 - \frac{2{,}445605}{2^{\frac{3}{2}}} \approx 69{,}01035$ | $81{,}875 + \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = 81{,}875 + \frac{2{,}445605}{2^{\frac{3}{2}}} \approx 82{,}73965$ |
| Syddanmark | 81 | $81 - \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = -\frac{2{,}445605}{2^{\frac{3}{2}}} + 81 \approx 80{,}13535$ | $81 + \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = \frac{2{,}445605}{2^{\frac{3}{2}}} + 81 \approx 81{,}86465$ |
| Midtjylland | 70 | $70 - \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = -\frac{2{,}445605}{2^{\frac{3}{2}}} + 70 \approx 69{,}13535$ | $70 + \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = \frac{2{,}445605}{2^{\frac{3}{2}}} + 70 \approx 70{,}86465$ |
| Nordjylland | 71,625 | $71{,}625 - \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = 71{,}625 - \frac{2{,}445605}{2^{\frac{3}{2}}} \approx 70{,}76035$ | $71{,}625 + \left(\frac{\sqrt[2]{2}}{2}\right) \cdot 2{,}63 \cdot \frac{\sqrt[2]{70{,}04}}{18} = 71{,}625 + \frac{2{,}445605}{2^{\frac{3}{2}}} \approx 72{,}48965$ |
| (95%,12) | | | |

Vi kan her til sidst se, at selve Region Hovedstad og Region Syddanmark er dem som har den højeste interval hvilket gør at de ligger øver i listen i forhold til Scoring. Hvorimod Region Sjælland ligger også godt øverst. Kigger man over imod Region Midtjylland og Nordjylland har de scoret mindst. Vi kan dermed se, at Middelværdierne for alle Regionerne er anderledes, som betyder at null-hypotesen er afvist men

# Statistisk Dataanalyse

Hovedstad, Sjælland og Syddanmark ligger øverst i Scoren og Midtjylland og Nordjylland ligger nederst i Scoren.