BI1363 HT 2020
Variansanalys
Oktober 2020
Adam Flöhr, BT, SLU

Variansanalys

Motsvarar *Biometri*, kap 10

I korthet

t-testet kan testa om två stickprov av kontinuerlig data har samma medelvärde

Naturlig utveckling att vilja testa **fler än två** stickprov

Möjligt genom **variansanalys** (eller ANalysis Of VAriance, *Anova*)

Genom variansanalys kan man testa om det finns *någon* skillnad mellan grupper genom ett **F-test**

Man kan även testa om *två* valda grupper är skilda med hjälp av **LSD (Least Significant Difference)**

Exempeldata

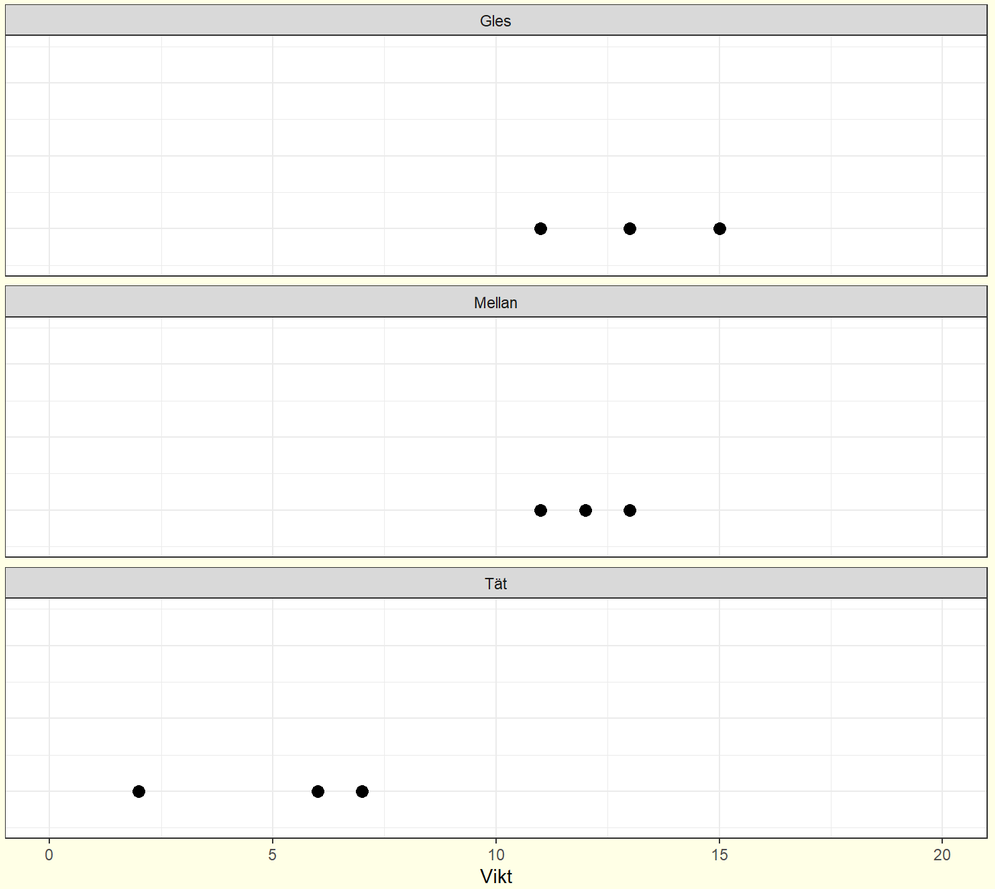
Testar en ärtgrödas reaktion på planteringstäthet.

Tre behandlingar: tät, mellan och gles

Utfallsvariabeln $\boldsymbol{y}$ är torrvikt i kilo per skörderuta

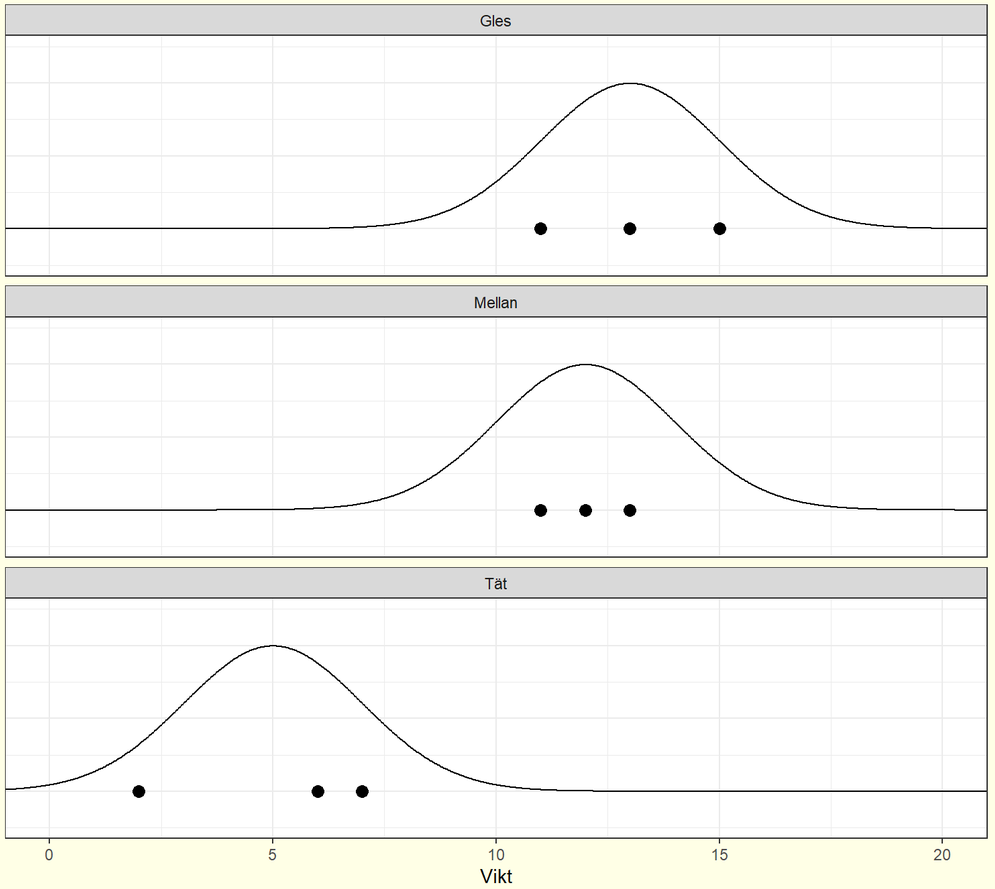
Behandling	**y**
Tät	2
Tät	7
Tät	6
Mellan	12
Mellan	13
Mellan	11
Gles	11
Gles	15
Gles	13

Illustration



Varje behandling innefattar tre observationer

Illustration



I en anova-modell är observationerna normalfördelade kring ett gruppspecifikt medelvärde

Det gruppspecifika medelvärdet skattas med det observerade medelvärdet för gruppen

Modellekvation

En observation är summan av

- ett övergripande medelvärde μ
- en behandlingseffekt α
- en slumpmässig felterm e

Detta kan uttryckas

$$y_{ij} = \mu + \alpha_i + e_{ij} \qquad e_{ij} \sim N(0, \sigma_e^2)$$

$$i = 1, \ldots, a \qquad j = 1, \ldots, n$$

Anova-tabell

Egenskaper hos en skattad anova-modell sammanfattas ofta i en anova-tabell

Anova-tabellen ger en uppdelning av variationen i en del *mellan* grupper och en del *inom* grupper

Kvadratsummor

Variansanalys bygger på beräknade kvadratsummor

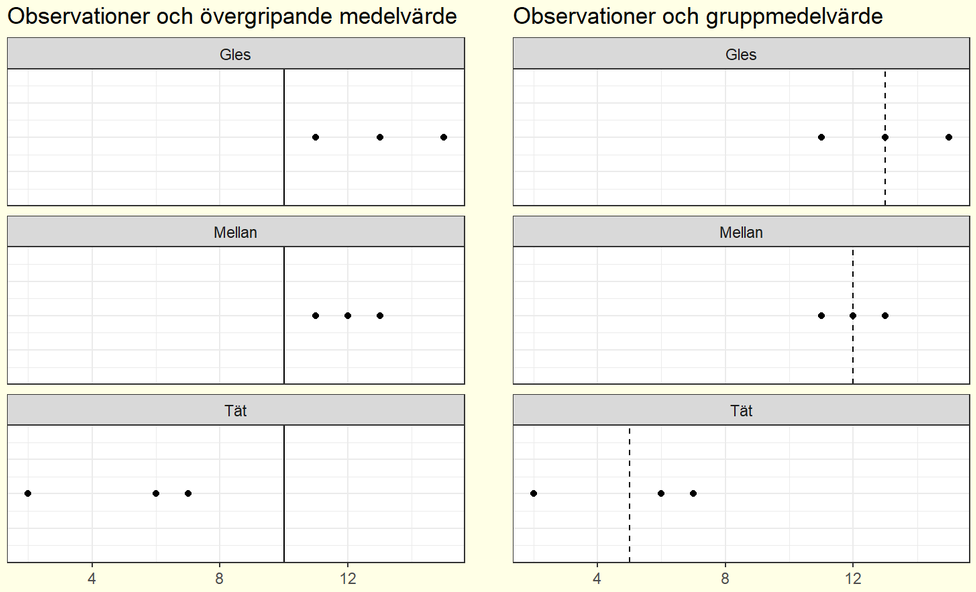
Datan ger skattningar för ett övergripande medelvärde och ett gruppmedelvärde

Summan av kvadrerade avstånd mellan observationen och det övergripande medelvärdet ger en kvadratsumma för totalen SS_T

Summan av kvadrerade avstånd mellan observationen och gruppmedelvärdet ger en kvadratsumma för slumpfelet SS_e

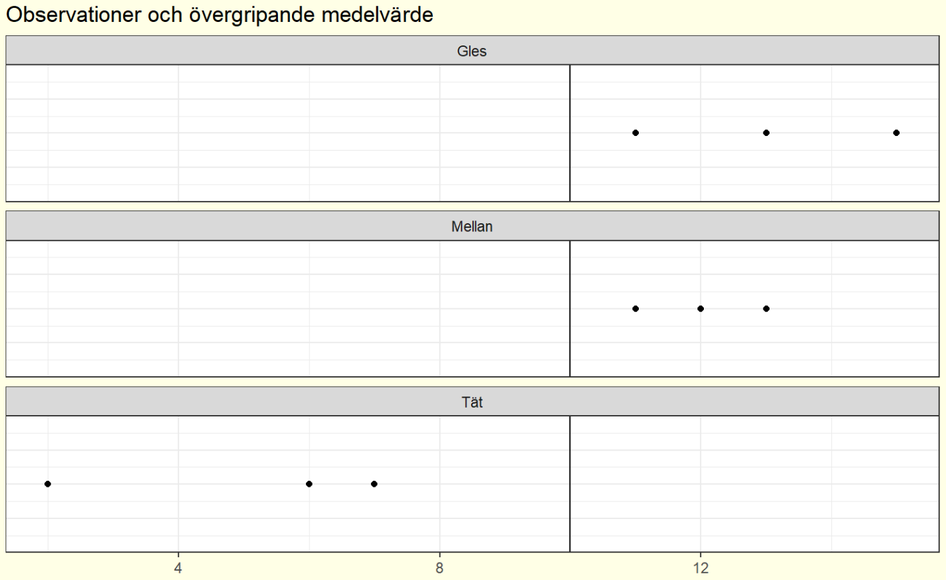
Differensen mellan SS_T och SS_e betecknas SS_A

$$SS_A = SS_T - SS_e$$



Kvadratsummor, exempel, SS_T

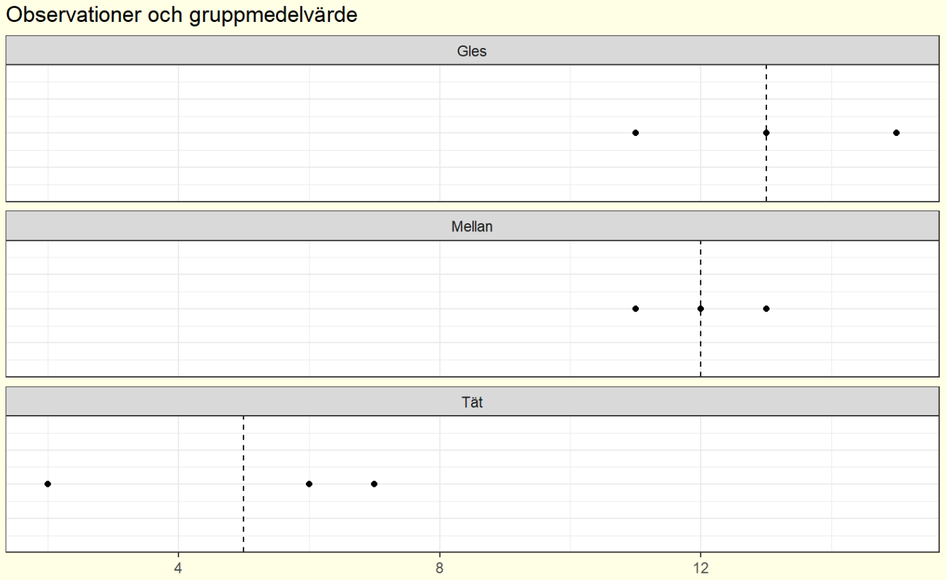
Behandling	Utfall	Övergripande medel	Residual	Kvadrerad residual
Tät	2	10	-8	64
Tät	7	10	-3	9
Tät	6	10	-4	16
Mellan	12	10	2	4
Mellan	13	10	3	9
Mellan	11	10	1	1
Gles	11	10	1	1
Gles	15	10	5	25
Gles	13	10	3	9



SS_T är summan av kvadrerade residualer $SS_T = 138$

Kvadratsummor, exempel, SS_e

Behandling	Utfall	Gruppmedel	Residual	Kvadrerad residual
Tät	2	5	-3	9
Tät	7	5	2	4
Tät	6	5	1	1
Mellan	12	12	0	0
Mellan	13	12	1	1
Mellan	11	12	-1	1
Gles	11	13	-2	4
Gles	15	13	2	4
Gles	13	13	0	0



SS_T är summan av kvadrerade residualer $SS_e = 24$

Frihetsgrader

Antalet frihetsgrader för en kvadratsumma ges av antalet observationer minus antalet skattade parametrar

Vid beräkningen av den totala kvadratsumman SS_T används en skattad parameter. antalet frihetsgrader är alltså $N - 1$

Vid beräkningen av kvadratsumman för slumpfelet skattas en parameter per grupp. För a stycken grupper ges antalet frihetsgrader därmed av $N - a$

Skillnaden i antalet frihetsgrader är antalet frihetsgrader som tillskrivs SS_A: $(N - 1) - (N - a) = a - 1$, det vill säga antalet grupper minus ett

Anova-tabell

Modellens resultat sammanfattas i en *Anova-tabell*

Anova-tabellen innehåller information om kvadratsummor (SS) och frihetsgrader (fg), utifrån dessa beräknas medelkvadratsummor (MS), F-värden, och p-värden

Generellt schema

	Df	Sum Sq (SS)	Mean Sq (MS)	F	p-value
Behandling (A)	$a-1$	SS_A	$SS_A/(a-1)$	MS_A/MS_e	
Residual (e)	$N-a$	SS_e	$SS_e/(N-a)$		
Total (T)	$N-1$	SS_T	$SS_T/(N-1)$		

Anova-tabell, exempel

I vårt exempel har vi tre behandlingar och nio observationer ($N = 9$ och $a = 3$)

	Df	Sum Sq (SS)	Mean Sq (MS)	F	p-value
Behandling (A)	$3 - 1$	SS_A	$SS_A/(a-1)$	MS_A/MS_e	
Residual (e)	$9 - 3$	SS_e	$SS_e/(N-a)$		
Total (T)	$9 - 1$	SS_T	$SS_T/(N-1)$		

Vi har tidigare beräknat kvadratsummorna $SS_T = 138$ och $SS_e = 24$.

SS_A ges av differensen $138 - 24 = 114$.

	Df	Sum Sq (SS)	Mean Sq (MS)	F	p-value
Behandling (A)	2	114	$SS_A/(a-1)$	MS_A/MS_E	
Residual (e)	6	24	$SS_e/(N-a)$		
Total (T)	8	138	$SS_T/(N-1)$		

Medelkvadratsummorna MS ges av kvadratsumman delat på antalet frihetsgrader

	Df	Sum Sq (SS)	Mean Sq (MS)	F	p-value
Behandling (A)	2	114	57	MS_A/MS_e	
Residual (e)	6	24	4		
Total (T)	8	138	17.25		

F-test

Har alla grupper samma populationsmedelvärde?

Testas med ett F-test

Samma gång som tidigare hypotestest (HTTPS)

Testfunktionen hämtas från anova-tabellen

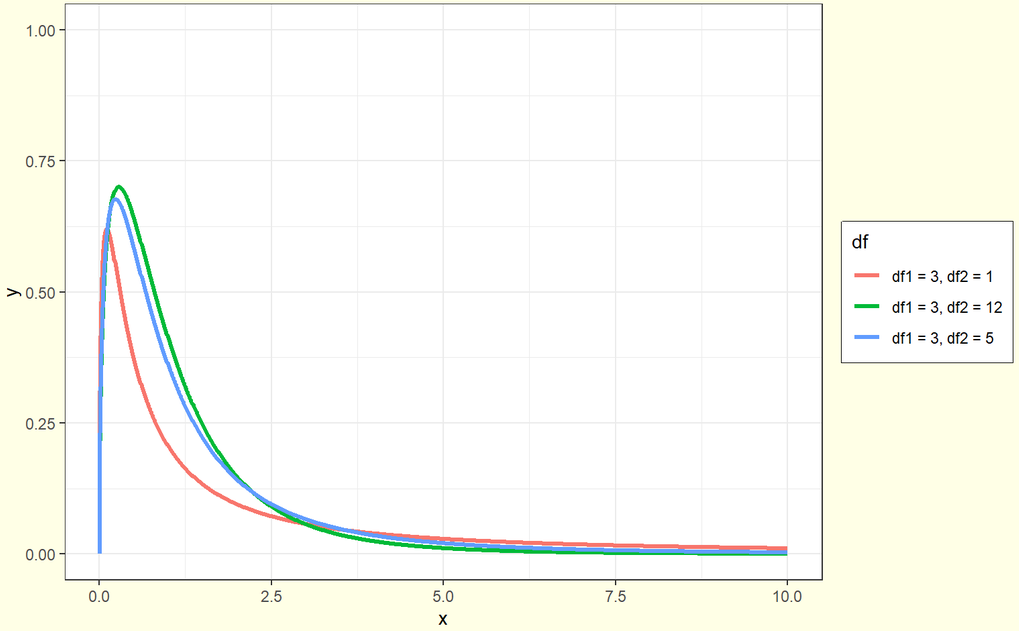
Testfördelningen är en F-fördelning

F-fördelning

Tester av kvadratsummor baseras på en testfördelning som kallas en F-fördelning

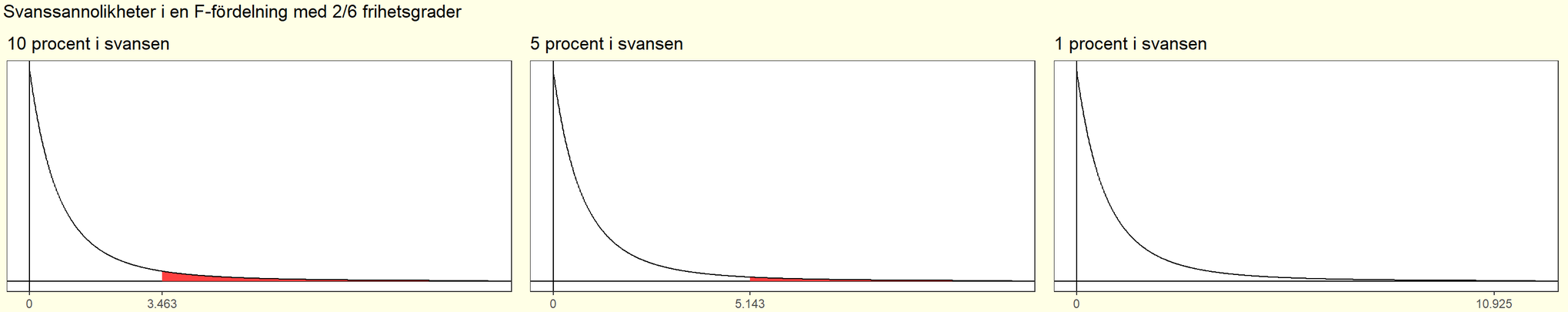
En F-fördelning uppstår som kvoten av två χ^2-fördelade slumpvariabler

Den defineras av två parametrar, antalet *frihetsgrader i täljaren* ν_1, och antalet *frihetsgrader i nämnaren* ν_2



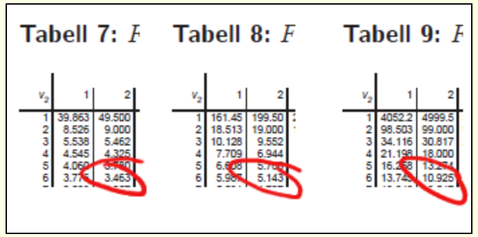
Biometri, tabell 8 (och 7, 9, 10)

Som tidigare kommer vi vilja uppskatta svanssannolikheten



Tabellvärden för x-axeln betecknas $F_{(1-\alpha,\nu_1,\nu_2)}$ och kan hämtas från en tabell över F-fördelningen

F-fördelningen delas i flera tabeller, beroende på värdet för α



Kolumnen ges av ν_1 och raden av ν_2

För ett tabellvärde motsvarande fem procent i svansen och frihetsgrader 2 och 6, tittar vi i tabell 8 för $F_{(0.95,2,6)} = 5.143$

F-test, schema

Hypoteser

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_a$ alla grupper har samma medelvärde

H_1 : det finns skillnad i medelvärde mellan några grupper

Testfunktion

$$F = \frac{MS_A}{MS_e}$$

där MS_A och MS_e hämtas från anova-tabellen

Testfördelning

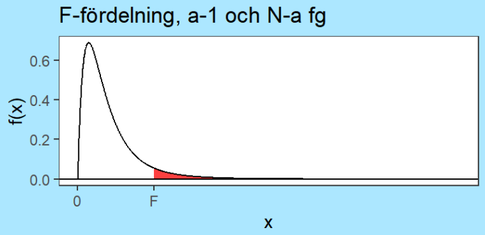
Under nollhypotesen följer F en F-fördelning med ν_1 och ν_2 frihetsgrader

I den enkla anova-modellen gäller $\nu_1 = a - 1$ och $\nu_2 = N - a$

P-värde

P-värdet ges av arean bortom F i testfördelningen

Vid handräkning uppskattas p-värdet genom att ställa F mot ett tabellvärde



Svar

P-värdet ställs mot en förbestämd *signifikansnivå* (ofta 5 procent)

Vid ett lågt p-värde förkastas nollhypotesen

Vid ett högt p-värde förkastas ej nollhypotesen

F-test, exempel

Vi beräknade tidigare följande anova-tabell

	Df	Sum Sq (SS)	Mean Sq (MS)	F	p-value
Behandling (A)	2	114	57	MS_A/MS_e	
Residual (e)	6	24	4		
Total (T)	8	138	17.25		

Hypoteser

H_0 : ingen skillnad mellan behandlingar

H_1 : någon skillnad mellan behandlingar

Testfunktion

$$F_{obs} = \frac{MS_A}{MS_e} = \frac{57}{4} = 14.25$$

MS_A och MS_e kan hämtas från anova-tabellen

Testfördelning

Under nollhypotesen följer $\boldsymbol{F}$ en F-fördelning med $\boldsymbol{\nu_1 = 2}$ och $\boldsymbol{\nu_2 = 6}$

P-värde

P-värdet ges av ytan till höger om vårt observerade $\boldsymbol{F = 14.25}$

Vi kan uppskatta p-värdet från *Biometri* tabell 6

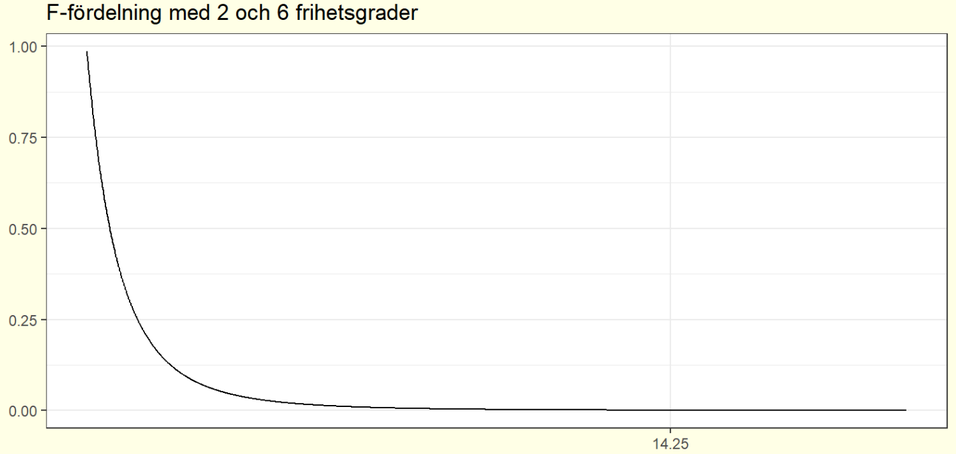
Vårt observerade värde ligger över $\boldsymbol{F_{(0.99,2,6)} = 10.925}$. P-värdet måste alltså vara under $\boldsymbol{1}$ procent

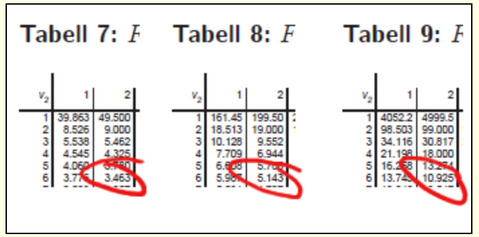
En datorberäkning ger det exakta värdet $\boldsymbol{0.00526}$

Slutsats

P-värdet är under $\boldsymbol{5}$ procent, vilket ger att vi förkastar nollhypotesen

Det finns någon statistiskt säkerställd skillnad i populationsmedelvärde mellan behandlingarna

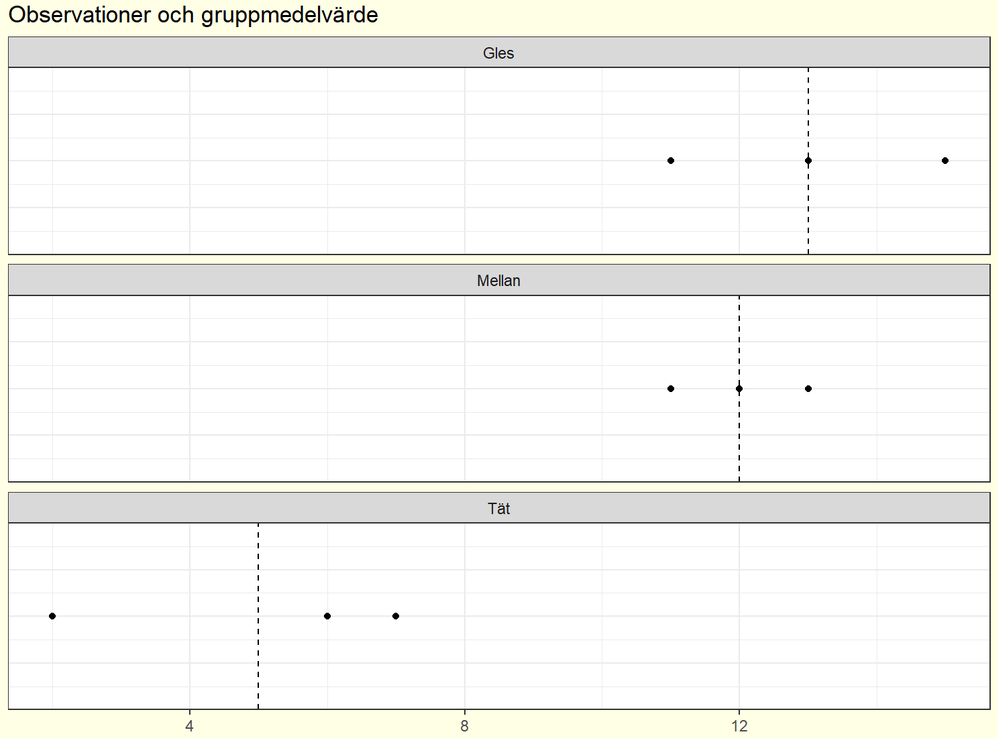




Parvisa jämförelser

Signifikans i F-testet ger att det finns *någon* skillnad mellan behandlingar

Vari ligger skillnaden?



Minsta signifikanta skillnad

Ett konfidensintervall för skillnaden mellan två gruppmedelvärden ges av

$$\bar{x_1} - \bar{x_2} \pm t_{(0.975,df_e)} \sqrt{MS_e(\frac{1}{n_1} + \frac{1}{n_2})}$$

Den sista termen kallas minsta signifikanta skillnad (*(Fisher's) Least Significant Difference, LSD*)

$$LSD = t_{(0.975,df_e)} \sqrt{MS_e(\frac{1}{n_1} + \frac{1}{n_2})}$$

Om skillnaden mellan två gruppers medelvärden överstiger LSD är grupperna signifikant skilda på fem-procentsnivån

Minsta signifikanta skillnad, exempel

I vårt exempel är $n = 3$ och medelvärden ges av

Behandling	Mean
Gles	13
Mellan	12
Tät	5

	Df	Sum Sq (SS)	Mean Sq (MS)	F	p-value
Behandling	2	114	57	14.25	0.005
Residual	6	24	4		
Total	8	138	17.25		

Vi beräknar LSD genom

$$LSD = t_{(0.975, df_e)} \sqrt{MS_e (\frac{1}{n_1} + \frac{1}{n_2})} = 2.447 \sqrt{4 (\frac{1}{3} + \frac{1}{3})} = 3.996$$

Behandling *Tät* är signifikant skild från *Mellan* och *Gles*, eftersom skillnaden i medelvärde överstiger 4

Behandlingarna *Mellan* och *Gles* är inte signifikant skilda, eftersom medelvärdesskillnaden är mindre än 4

Signifikansbokstäver

Parvisa jämförelser presenteras ofta med signifikansbokstäver

Behandlingar som *inte* är signifikant åtskilda delar någon bokstav

I vårt enkla fall får vi

Behandling	**Medel**	**Sig-bokstäver**
Gles	13	a
Mellan	12	a
Tät	5	b

C och B är inte signifikant åtskilda och delar därmed en bokstav *a*

A är signifikant skild från B och C och delar därmed ingen bokstav

Mass-signifikans

Ett tests signifikansnivå anger sannolikheten att förkasta nollhypotesen även om den stämmer

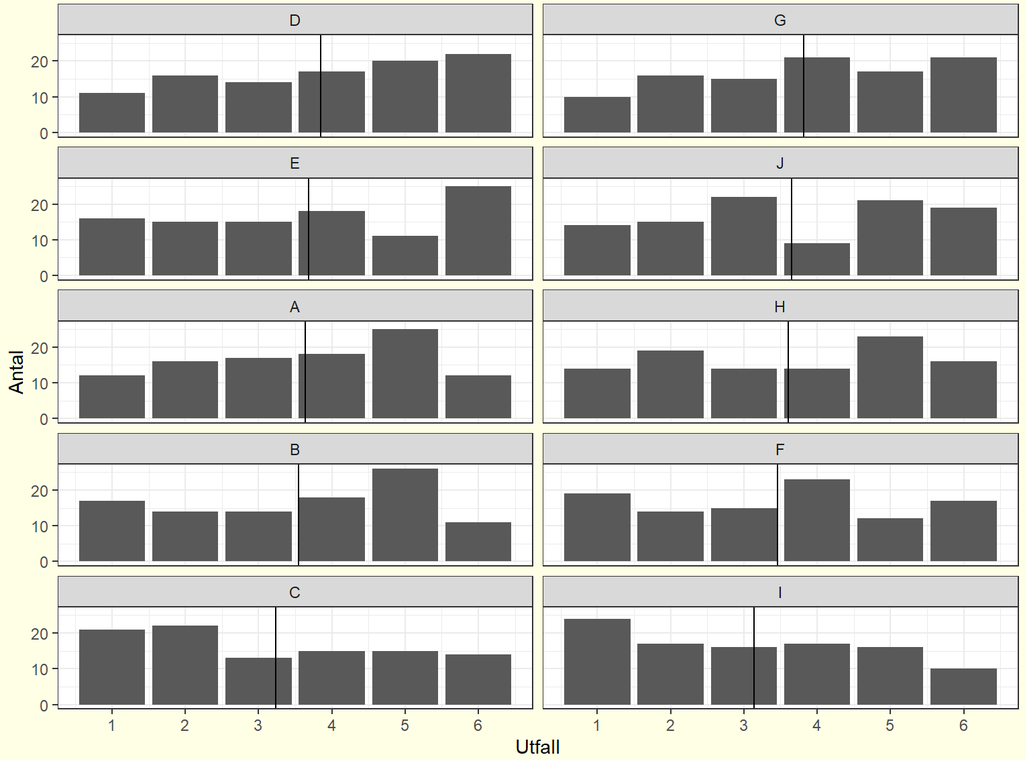
Om vi jämför två behandlingar på fem-procentsnivån finns alltså en fem-procentig risk att få ett signifikant resultat även om behandlingarna är exakt lika

Vid parvisa jämförelser utför man ett test per par. Risken att få *något* signifikant resultat kan därmed bli väldigt stor även om alla behandlingar är lika

Fenomenet kallas för *mass-signifikans*

Protected LSD. Gör bara parvisa jämförelser om F-testet visar på någon signifikant skillnad

Honest significant difference. Justera LSD-beräkningen



Honest Significant Difference (HSD)

För att kompensera för mass-signifikans används ofta *Honest Significant Difference (HSD)* istället för LSD

Även kallat Tukey-test eller Tukey-metoden

HSD har samma tolkning som LSD: en medelvärdesskillnad större än HSD innebär att grupperna är statistiskt skilda

HSD är i regel större än LSD och leder därmed till färre signifikanta skillnader

Blockförsök

I ett försök har man objekt som man mäter (*försöksenheter*) och *behandlingar* man utsätter dem för

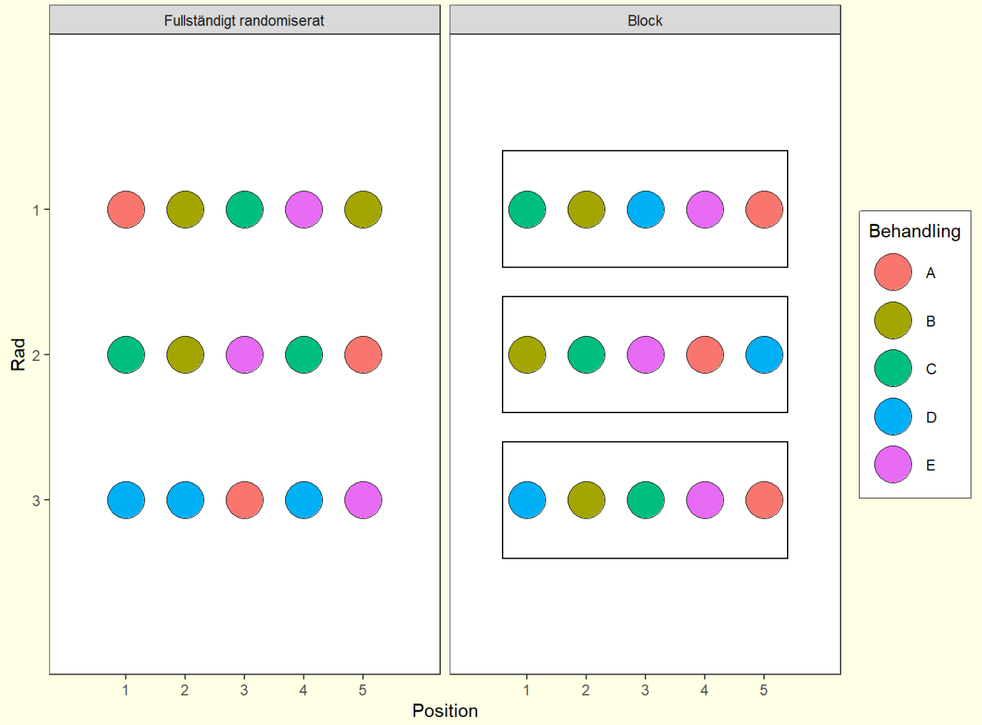
Varje försöksenhet tilldelas en behandling

Fullständigt randomiserad design

- Varje behandling förekommer ett givet antal gånger
- I övrigt tilldelas enheter behandling slumpmässigt
- Alla mönster är möjliga

Blockdesign

- Försöksenheterna delas in i undergrupper (*block*)
- Varje behandling förekommer en gång i varje block



Tanken med block

Målet med en blockdesign är att gruppera *lika* försöksenheter

Eftersom varje behandling förekommer i varje block, får man en rättvis jämförelse mellan lika enheter

Växthus

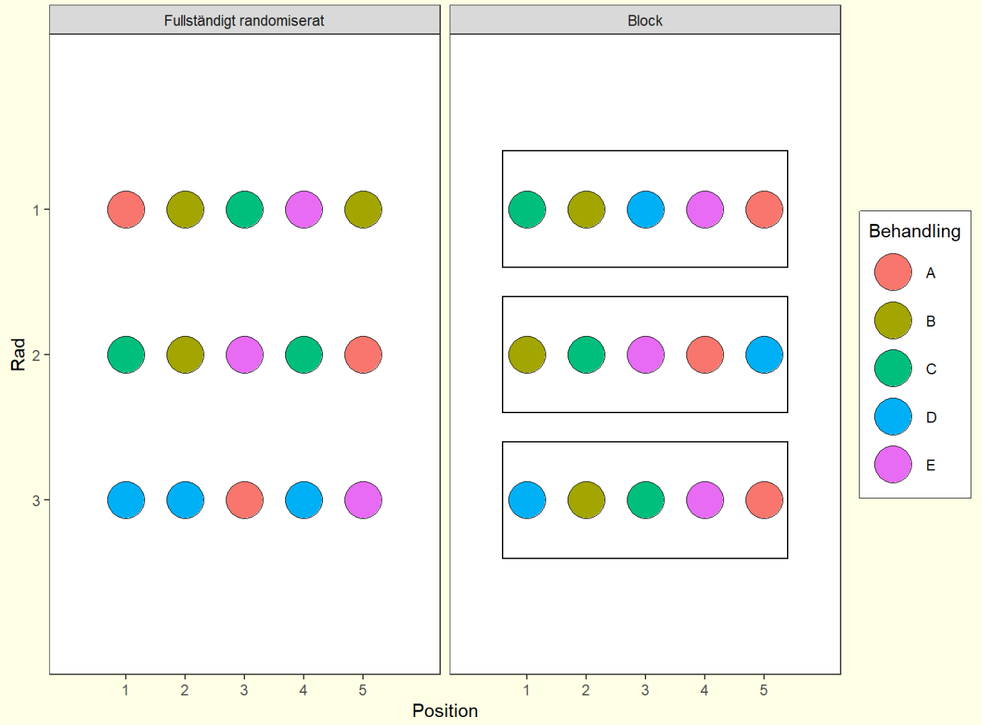
Fem näringsbehandlingar, tre plan i en hylla

Eftersom nivå kan påverka resultatet sätter man varje behandling på varje hyllplan i en blockdesign

Restaurang

Fem restauranger, tre måltidsdelar (förrätt, huvudrätt, efterrätt)

För en rättvis jämförelse beställer man en förrätt, en huvudrätt och en efterrätt från varje restaurang



Anova med block

Ett blockförsök analyseras med *variansanalys med block*

Modellekvation

En observation är summan av

- ett övergripande medelvärde μ
- en behandlingseffekt α
- en blockeffekt β
- en slumpmässig felterm e

Detta kan uttryckas

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \qquad e_{ij} \sim N(0, \sigma_e^2)$$

$$i = 1, \ldots, a \qquad j = 1, \ldots, n$$

Anovatabell med block

Eftersom vi har en tillagd term får vi en ny rad i anova-tabellen

Det generella schemat blir

	Df	Sum Sq (SS)	Mean Sq (MS)	F	p-value
Block (Bl)	$b-1$	SS_Bl	$SS_{Bl}/(b-1)$	MS_{Bl}/MS_e	
Behandling (A)	$a-1$	SS_A	$SS_A/(a-1)$	MS_A/MS_e	
Residual (e)	$N-a-b+1$	SS_e	$SS_e/(N-a)$		
Total (T)	$N-1$	SS_T	$SS_T/(N-1)$		

I ett balanserat försök är $N = a \cdot b$, vilket ger de alternativa uttrycken $df_T = ab - 1$ och $df_e = (a-1)(b-1)$

Under nollhypotesen att det inte finns någon behandlingseffekt kommer MS_A/MS_E från en F-fördelning med $a-1$ och $N-a-b+1$ frihetsgrader

Under nollhypotesen att det inte finns någon blockeffekt kommer MS_{Bl}/MS_E från en F-fördelning med $b-1$ och $N-a-b+1$ frihetsgrader

Anova med block, exempel

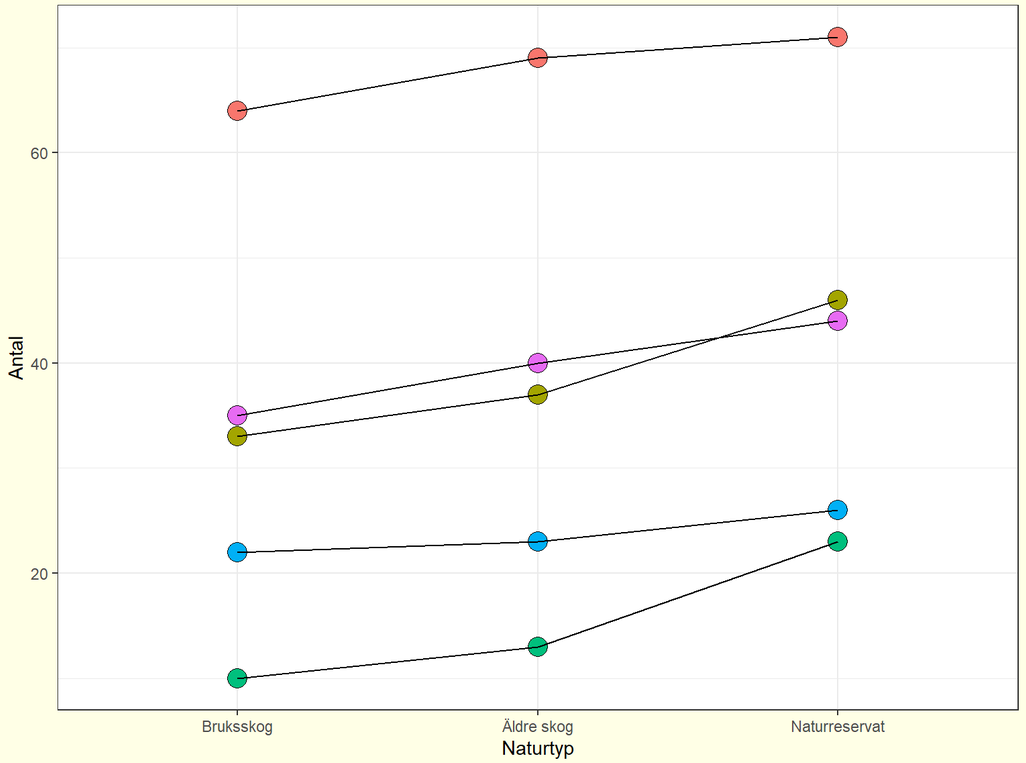
Vi vill undersöka populationsstorleken på smalbandad ekbarrbock (*Plagianotus Arcuatus*)

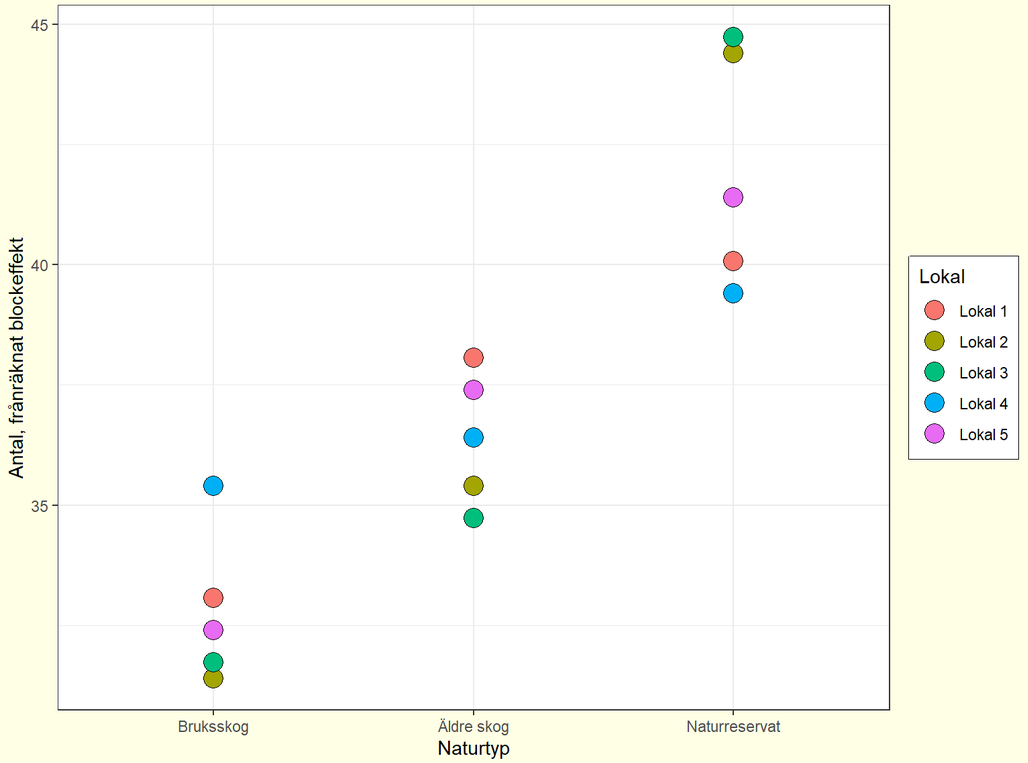
Sätter upp fällor i tre typer av områden (bruksskog, äldre skog, och naturreservat)

För en rättvis jämförelse väljer vi ut fem lokaler där vi kan hitta alla tre naturtyper nära varandra

Det här ger en blockdesign med naturtypen som behandling och lokalen som block

Lokal	Naturtyp	Antal
Lokal 1	Bruksskog	64
Lokal 1	Äldre skog	69
Lokal 1	Naturreservat	71
Lokal 2	Bruksskog	33
Lokal 2	Äldre skog	37
Lokal 2	Naturreservat	46
Lokal 3	Bruksskog	10
Lokal 3	Äldre skog	13
Lokal 3	Naturreservat	23
Lokal 4	Bruksskog	22
Lokal 4	Äldre skog	23
Lokal 4	Naturreservat	26
Lokal 5	Bruksskog	35
Lokal 5	Äldre skog	40
Lokal 5	Naturreservat	44





Om man inte tar med blockdesignen överlappar observationer från olika naturtyper varandra

Med blocken frånräknade finns en klarare separation av behandlingar

Anova-tabell med block

Vi skattar en modell med block och får följande anovatabell

Typ	Df	Sum Sq (SS)	Mean Sq (MS)	F
Lokal	4	4854.3	1213.6	232.6
Naturtyp	2	214.9	107.5	20.6
Residual	8	41.7	5.2	
Total	14	5110.9	365.1	

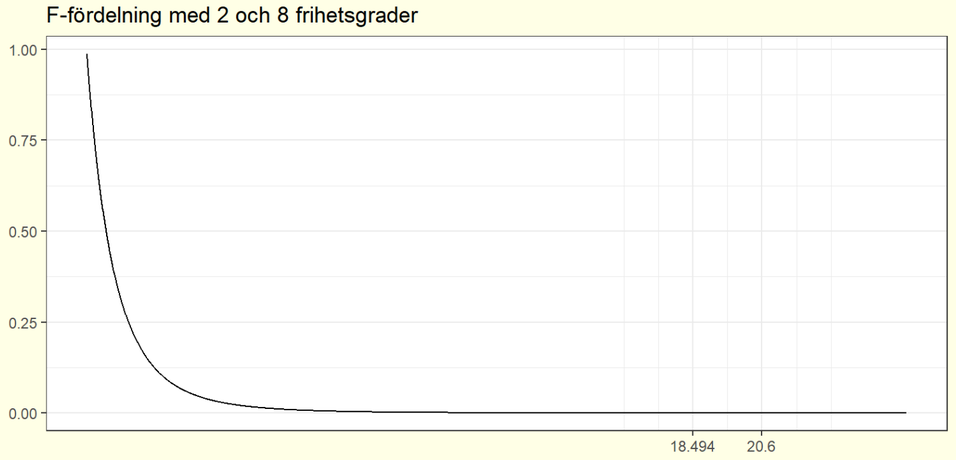
	Df	Sum Sq (SS)	Mean Sq (MS)	F	p-value
Block (Bl)	$b-1$	SS_{Bl}	$SS_{Bl}/(b-1)$	MS_{Bl}/MS_e	
Behandling (A)	$a-1$	SS_A	$SS_A/(a-1)$	MS_A/MS_e	
Residual (e)	$N-a-b+1$	SS_e	$SS_e/(N-a)$		
Total (T)	$N-1$	SS_T	$SS_T/(N-1)$		

Vi kan testa nollhypotesen att det inte finns någon skillnad mellan naturtyper

Under nollhypotesen kommer F-värdet för naturtyp ($\boldsymbol{F = 20.6}$) från en F-fördelning med $\boldsymbol{\nu_1 = 2}$ och $\boldsymbol{\nu_2 = 8}$

Tabell 10 ger $\boldsymbol{F_{(0.999,2,8)} = 18.494}$. Vårt p-värde är alltså mindre än 0.001

Vi förkastar nollhypotesen och säger att det finns en statistiskt signifikant skillnad mellan naturtyper



LSD vid block

Behandlingar kan jämföras parvis på samma sätt som vid enkel variansanalys

Formeln för Fishers LSD ges av

$$LSD_{bl} = t_{(0.975,df_e)}\sqrt{MS_e\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = t_{(0.975,(a-1)(b-1))}\sqrt{MS_e \cdot \frac{2}{b}}$$

I vårt exempelfall får vi

$$LSD_{bl} = t_{(0.975,8)}\sqrt{5.2 \cdot \frac{2}{5}} = 2.306 \cdot 1.442 = 3.325$$

Typ	Df	Sum Sq (SS)	Mean Sq (MS)	F
Lokal	4	4854.3	1213.6	232.6
Naturtyp	2	214.9	107.5	20.6
Residual	8	41.7	5.2	
Total	14	5110.9	365.1	

Naturtyperna har medelvärdena

Naturtyp	Medel	Sig
Bruksskog	32.8	a
Äldre skog	36.4	b
Naturreservat	42.0	c

Eftersom varje differens är större än LSD är alla typer signifikant skilda från de övriga

Det ger att varje typ har en egen signifikansbokstav

Slut