PSY2014 – Seminar

# Seminar #1 – Grunnleggende kvantitative konsepter

## Sigma notasjon

Den greske bokstaven Sigma ΣΣ brukes i matematikken for å representerer en *sum*. Ved å bruke dette symbolet kan vi på effektiv måte vise til summen av mange tall, uten å måtte skrive mye tekst. En sum av tall inngår i utregningen av de fleste statistikker, så vi bruker anledningen først til å friske opp bruken av sigma notasjon.

Anta at X er følgende sett av tall; X={2,2,−2,0,5,−1}X={2,2,−2,0,5,−1}

1. Hvilket tall er gitt ved X4X4? Hva er X2+X5X2+X5?
   * X4 = 0
   * X2+X5=2+5=7
2. Regn ut Σ4i=2(Xi)Σi=24(Xi)
   * Starte på det andre elementet, summere til det fjerde elementet
   * i = hvilket element i settet du refererer til
   * 2+(-2)+0+5=5
   * X2+X3+X4
3. Regn ut Σni=1(Xi−1)2Σi=1n(Xi−1)2
   * Summere
   * n = alle elementene i settet siden n ikke er spesifisert
   * Ta alle elementene i settet og trekke fra 1
   * (2-1)^2 + (2-1)^2 + (-2-1)^2 + (0-1)^2 + (5-1)^2 + (-1-1)^2 = 1 + 1 + 9 +1 +16 + 4 = 32
4. Gjennomsnittsverdien av settet X er for lik 1.0 (typisk skrevet X¯=1.0X¯=1.0). Hvilken statistikk har du regnet ut dersom du multipliserer resultatet fra deloppgave 3 med 0.2?
   * 1 er gjennomsnittet, så på forrige oppgave så trakk vi fra gjennomsnittet fra alle elementene
   * 0.2 = 1/5
   * I varians deler vi på n-1
   * Derfor har vi regnet ut variansen

## Beskrivende statistikk

Er det et forhold mellom leseferdigheter og selvtillit hos barn? I tabellen under finner du mål på leseferdigheter og selvtillit hos fem åtteåringer.

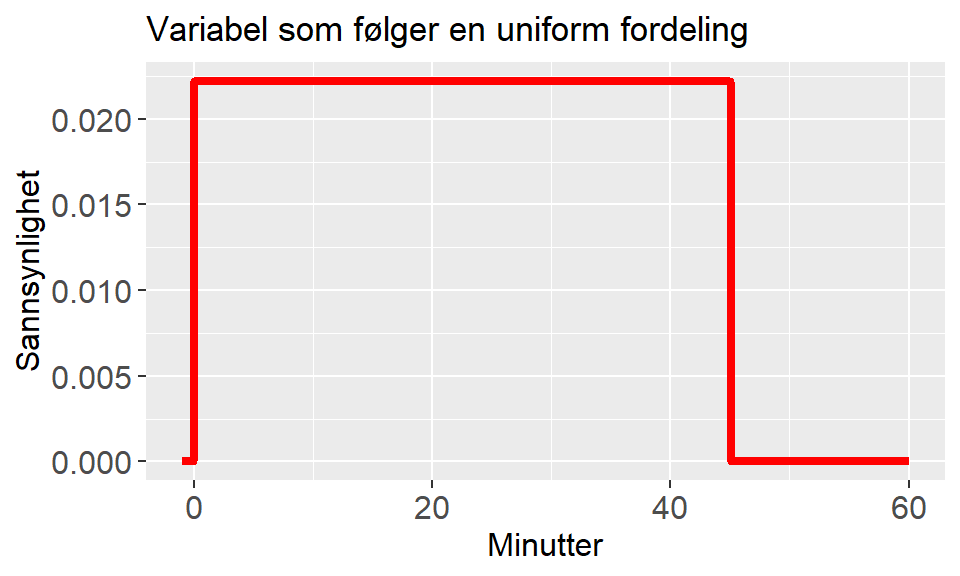
| ID | LESEFERDIGHET | SELVTILLIT |
| --- | --- | --- |
| 1 | 4 | 2 |
| 2 | 2 | 2 |
| 3 | 4 | 4 |
| 4 | 6 | 3 |
| 5 | 4 | 4 |

1. For enkelhets skyld vil vi kalle variabelen LESEFERDIGHET for XX. Regn ut *gjennomsnittet* (X¯X¯) og *variansen* (s2XsX2).
   * X med strek over betyr gjennomsnitt
   * Gjennomsnitt: 4
   * Varians:
2. Hva ville hatt størst påvirkning på gjennomsnittet og variansen til LESEFERDIGHET; at alle barna i utvalget økte sin leseferdighet med 5 poeng, eller at de doblet sin nåværende skåre? Prøv å svare på dette uten å regne ut de nye verdiene.
   * Varians: doble skåre
   * Gjennomsnitt: legge til 5
3. Variansen til SELVTILLIT er 1.0. Regn ut *kovariansen* og *korrelasjonen* mellom LESEFERIGHET og SELVTILLIT. Svar også på følgende spørsmål; “Dersom du kjente alle barnas leseferdighet, hvor mye av variansen i selvtillit kunne du gjøre rede for”? (det vi kaller forklart varians).
   * Varians:
   * Korralasjon:
   * Forklart varians:

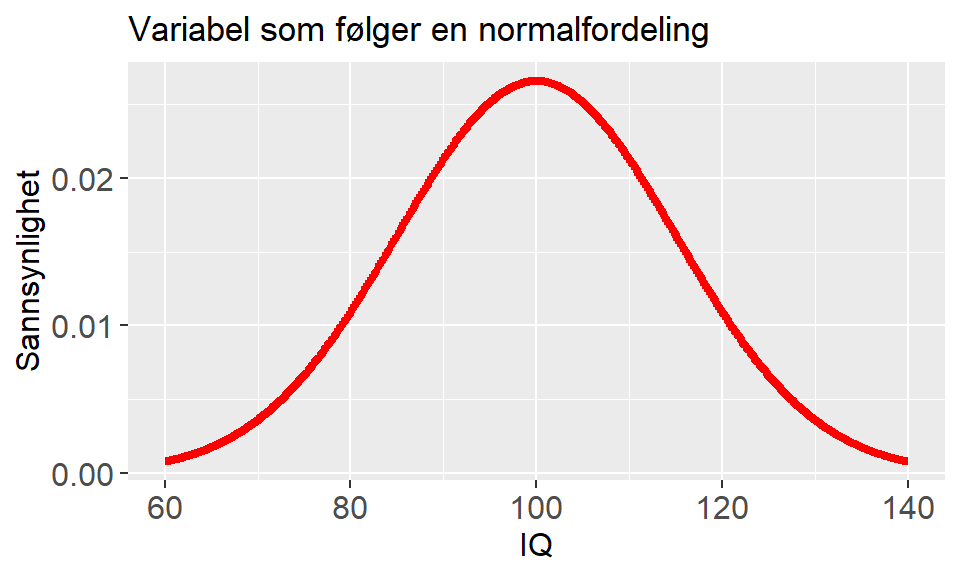
## Fordelinger

Vi vil møte en rekke fordelinger i PSY2014, så la oss nå få litt trening i hvordan slike fordelinger vanligvis brukes. La oss først betrakte en *uniformt fordelt variabel*, det vil si en variabel der sannsynlighetstettheten til alle mulige utfall er like stor.

Telefonen til Marit ringer en gang under hver forelesningstime (45 min). Vi vet ikke når den ringer, og alle tidspunkter er like sannsynlige. Vi kan si at den stokastiske variabelen T (tidspunktet der telefonen ringer) er uniformt fordelt på intervallet 0-45 minutter. Du finner en skisse av denne fordelingen under.

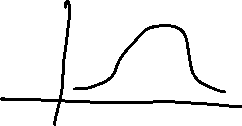


1. Hva er sannsynligheten for at telefonen ringer i løpet av første halvdel av forelesningen?
   * 50%
2. Hva er sannsynligheten for at telefonen ringer enten i løpet av de første 5 eller siste 5 minuttene?
   * 22%



I figuren over ser du et eksempel på en variabel som følger en *normalfordeling*. I befolkningen er IQ normalfordelt med snitt 100, og standardavvik 15. Normalfordelingen er en av de aller viktigste fordelingene innen statistikken, og vi vil etter hvert se at den beskriver mange fenomener svært godt. Sannsynligheten for at en normalfordelt variabel har et utfall i et gitt intervall kan ikke regnes på samme måte som i oppgaven over, men vi kan bruke tabeller til å finne dem. Dere finner en tabell for normalfordelingen bakerst i Agresti, som «TABLE A»).

1. Anta at du skal åpne en butikk. Du kan ikke vite hvor mange kunder du vil ha det første året, men basert på tall fra andre bedrifter i nabolaget tror du at antall kunder vil følge en normalfordeling med snitt 4000 og standardavvik 1000, (N(4000,1000)N(4000,1000)).
2. Lag en skisse av denne fordelingen.
   * klokkeformet



1. Hva er sannsynligheten for at du vil ha mellom 3000 og 5000 kunder?
   * 68%
2. Hvilket intervall rundt gjennomsnittet rommer 95% av mulige utfall?
   * 2000-6000 (95%)
3. Dersom du har mindre enn 2500 kunder det første året går butikken konkurs. Hva er sannsynligheten for konkurs? (Hint: Regn ut en Z-skåre og bruk TABLE A).
   * Z-skåre:
   * Sannsynligheten for å gå konkurs er 6,68%
4. Hvilket intervall rundt gjennomsnittet rommer 50% av mulige utfall (hint: Vis at den øvre grensen er omtrent ved en z-skåre på 0.67. Hvor mange kunder tilsvarer dette?)

# Seminar #2

## Bivariat regresjon

For å bli kjent med de ulike statistikkene som skrives ut når man kjører en regresjonsanalyse i R/SPSS skal vi starte med å regne ut noen av de viktigste for hånd (med kalkulator). Anta at du har dataene (X og Y) gitt i tabellen under. Y skal være avhengig variabel, mens X er uavhengig (forklarings) variabel.

| ID | Xi | Yi | Yi^ | (Yi−Yi^) | (Yi−Yi^)^2 | (Yi−Yi¯)^2 |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 1 | 0.5+1.5\*1=2 | 1-2=-1 | (-1)^2 |  |
| 2 | 2 | 3 | 0.5+1.5\*2=3.5 | 3-3.5=-0.5 | (-0.5)^2 |  |
| 3 | 6 | 11 | 0.5+1.5\*6=9.5 | 11-9.5=1.5 | 1.5^2 |  |
| 4 | 4 | 3 | 0.5+1.5\*4=6.5 | 3-6.5=-3.5 | (-3.5)^2 |  |
| 5 | 2 | 7 | 0.5+1.5\*2=3.5 | 7-3.5=3.5 | 3.5^2 |  |

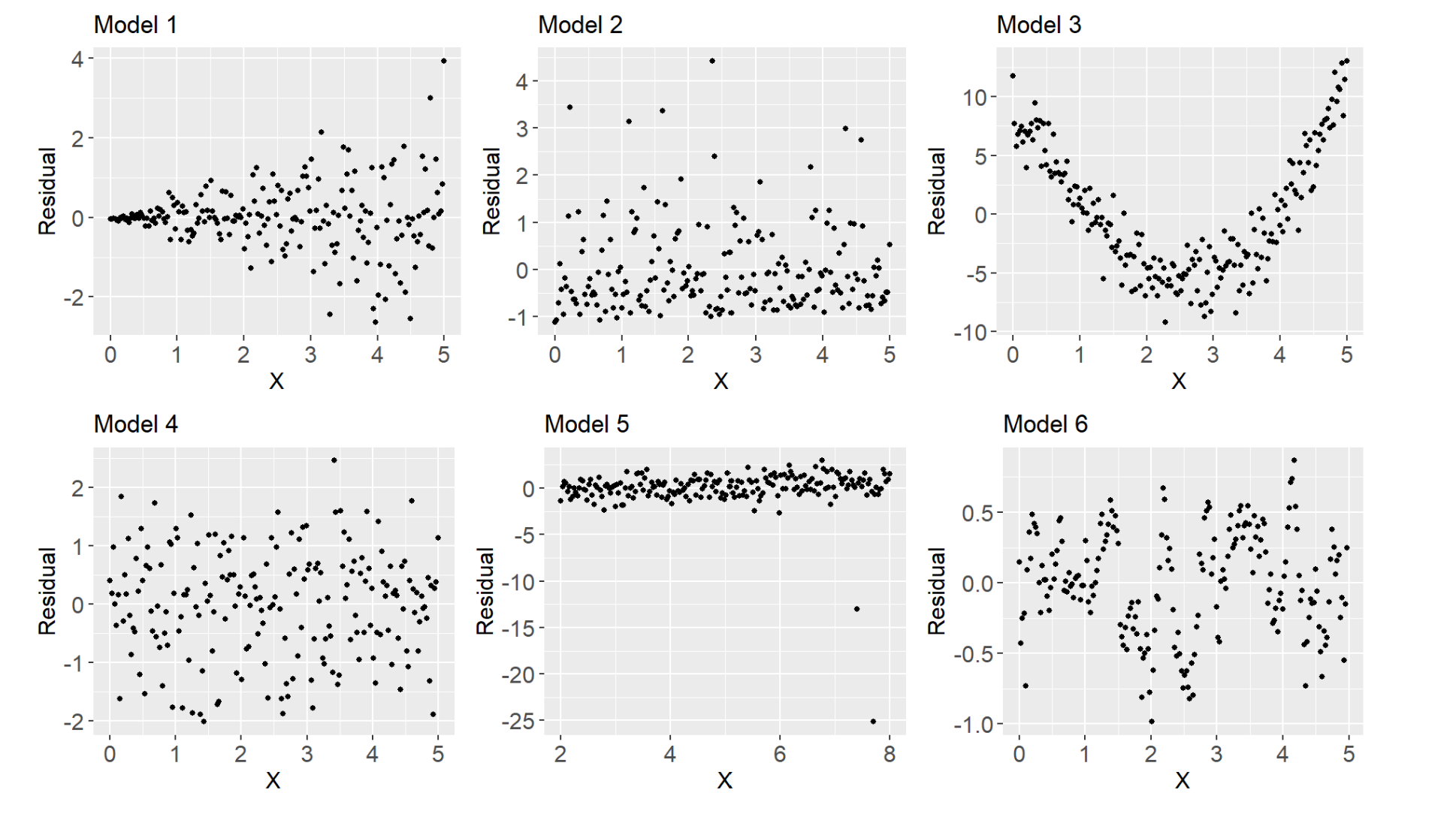
1. Tegn et spredningsplot (scatter plot) av X og Y. Før inn en linje som representerer din gjetning på hvor en regresjonslinje vil gå.
2. Anta at følgende statistikker er kjent. X¯=3X¯=3, sx2=4sx2=4, Y¯=5Y¯=5, sy2=16sy2=16. Bruk formlene gjennomgått på forelesning (normallikningene, som du også finner i avsnitt 9.2 i Agresti) for å finner minste kvadraters estimater for b0b0 og b1b1. Bruk disse estimatene til å skrive uttrykket for den forventede Y-verdien for en gitt X-verdi (prediction equation).
3. Hva er den forventede verdien (Yi^Yi^) og residualen (Yi−Yi^Yi−Yi^) for hver av de fem datapunktene i tabellen?
   * Se tabell
4. Hva er den forventede Y-verdien for en person med X-verdi lik 20 basert på din modell? Hvorfor burde du ikke stole på denne forventede verdien?
   * Vi burde ikke stole på denne forventede verdien fordi den går langt utenfor vår data. Ekstrapolasjon.
5. Bruk de estimerte verdiene for b0b0 og b1b1 til å tegne minste kvadraters regresjonslinje inn i spredningsplottet i oppgave 1. Skisser størrelsene Total sum of squares (TSS), Sum of Squares Model (SSM), og Sum of squared Error (SSE) i plottet, og regn ut verdiene.
6. Regn ut den forklarte variansen i Y (også kalt R2R2 eller “coefficient of determination”). Bruk denne verdien til å finne Perason korrelasjonen mellom X og Y (Hint: Husk hvordan vi brukte korrelasjonen til å finne forklart varians i oppgavesett 1).
7. Finn den *standardiserte regresjonskoeffisienten* for X, (β1^).
8. Hva uttrykker standardiserte regresjonskoeffisienten?
   * Den standardiserte regresjonskoeffisienten forteller oss om forventet antall standardavviksendring i y for hvert standardavvik i X. samme tolkning som r, men oppgis med ‘antall standardavvik’ som enhet.
9. Hvor mye forventer du at Y endrer seg når X endrer seg en mengde som tilsvarer:
   * Ett standardavvik
     1. X endres med ett standardavvik, Y endres med 0.75
   * To standardavvik
     1. For X, Y vil endres 2\*0.75=1.5
   * Verdien 3 (Tips: Hvor mange standardavvik tilsvarer dette?)
     1. Tilsvarer 1.5 standardavvik for X. 1.5\*0.75=1.125 for Y.

# Seminar #3

## 1. Brudd på antagelser i lineær regresjon

En enkel lineær regresjonsanalyse ble utført på 6 ulike datasett, og i fem av datasettene er en sentral antagelse ved lineær regresjon brutt. Prøv å identifisere hvilken av de seks tilfellene under som korresponderer til de ulike spredningsplottene. (Disse viser residualene plottet mot verdien på den uavhengige variabelen).

1. Et ikke-lineært forhold mellom X og Y.
2. Ingen antagelser brutt.
3. Uteliggere med stor innflytelse.
4. Korrelerte residualer (ikke uavhengige observasjoner).
5. Ikke-konstant feilvarians.
6. Ikke-normalfordelt feilvarians.



Residualer: feilvarians

Hvilke antagelser er brutt?

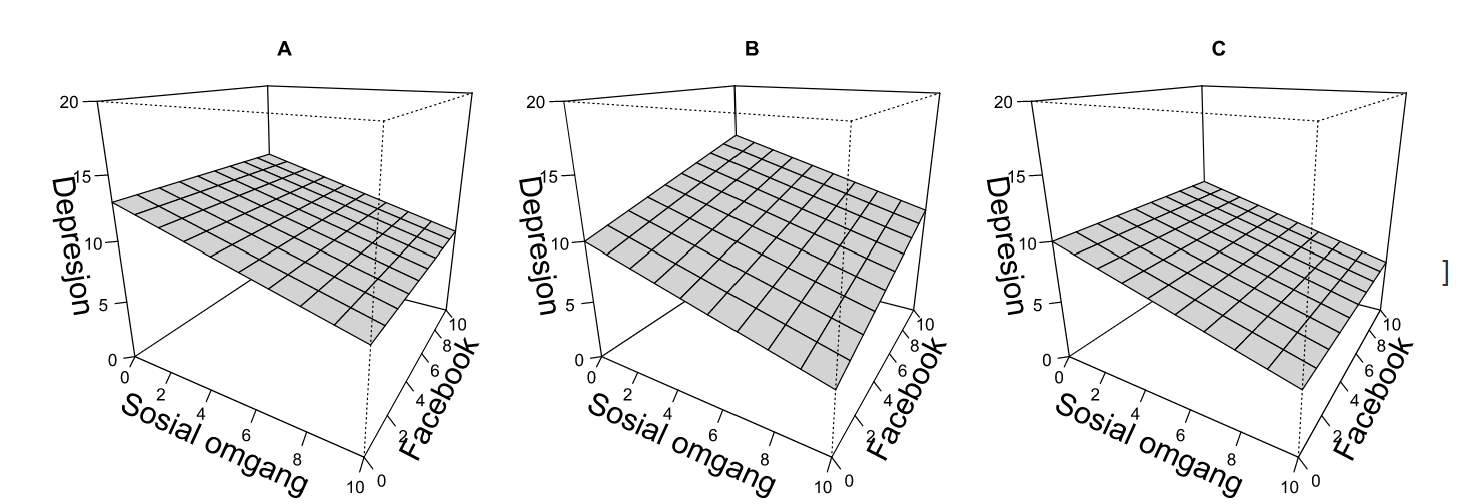
* + Modell 1: ikke-konstant feilvarians
  + Modell 2: ikke-normalfordelt feilvarians
  + Modell 3: ikke-lineært forhold mellom x og y
  + Modell 4: ingen, det er sånn vi vil det skal se ut når det er normalfordelt
  + Modell 5: uteliggere med stor innflytelse
  + Modell 6: korrelerte residualer. Klumper seg sammen på ulike x-verdier. Observasjonene våres kanskje ikke helt uavhengig.

## 2. flersvarsoppgaver

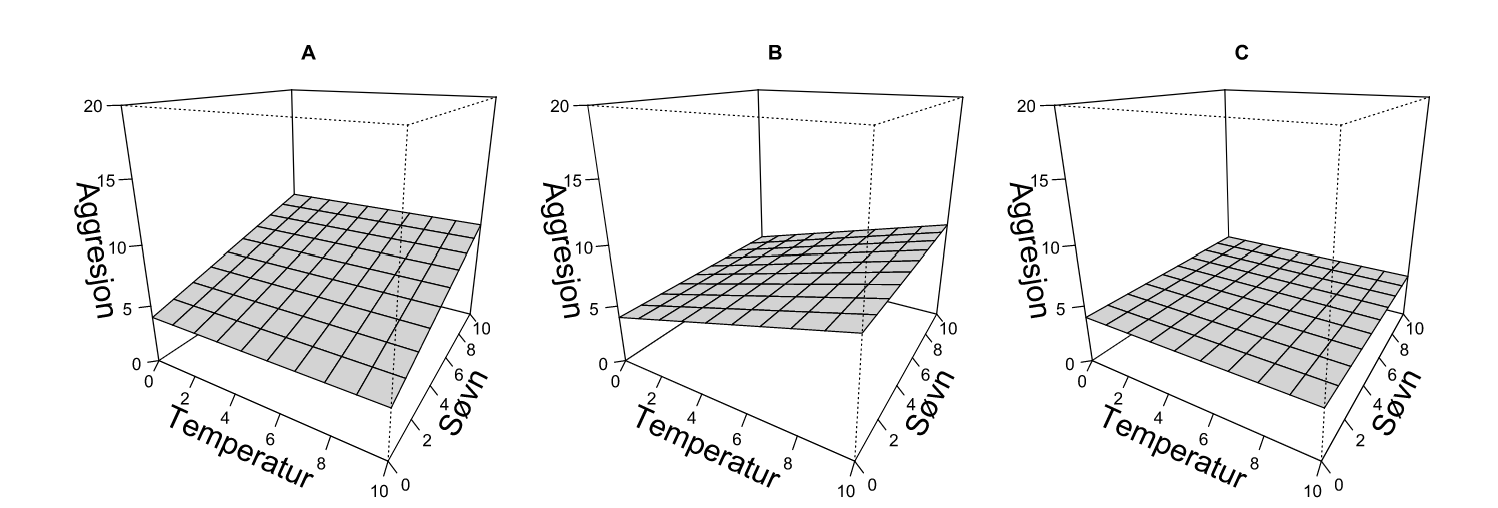
1. Dersom alle observasjonene i en multippel regresjonsmodell med to uavhengige variable faller rett på regresjonsplanet, hva blir R2?
   * forklart varians blir én. Om det er en perfekt korrelasjon, så er r=1, og r^2=1 også

* R^2= 1 - SSE/ESS. vi tar 1- fordi det er 100%, så om SSE/ESS, så forklarer modellen 100%

1. Dersom R2 er 0.975, hva kan du si om stigningstallet til regresjonslinjen?
   * den kan være enten positiv eller negativ fordi når man opphøyer noe i annen, så blir det positivt uansett hva du har opphøyd i annen enten det er et negativt eller positivt tall. -0.2^2 og 0.2^2 blir begge 0.4

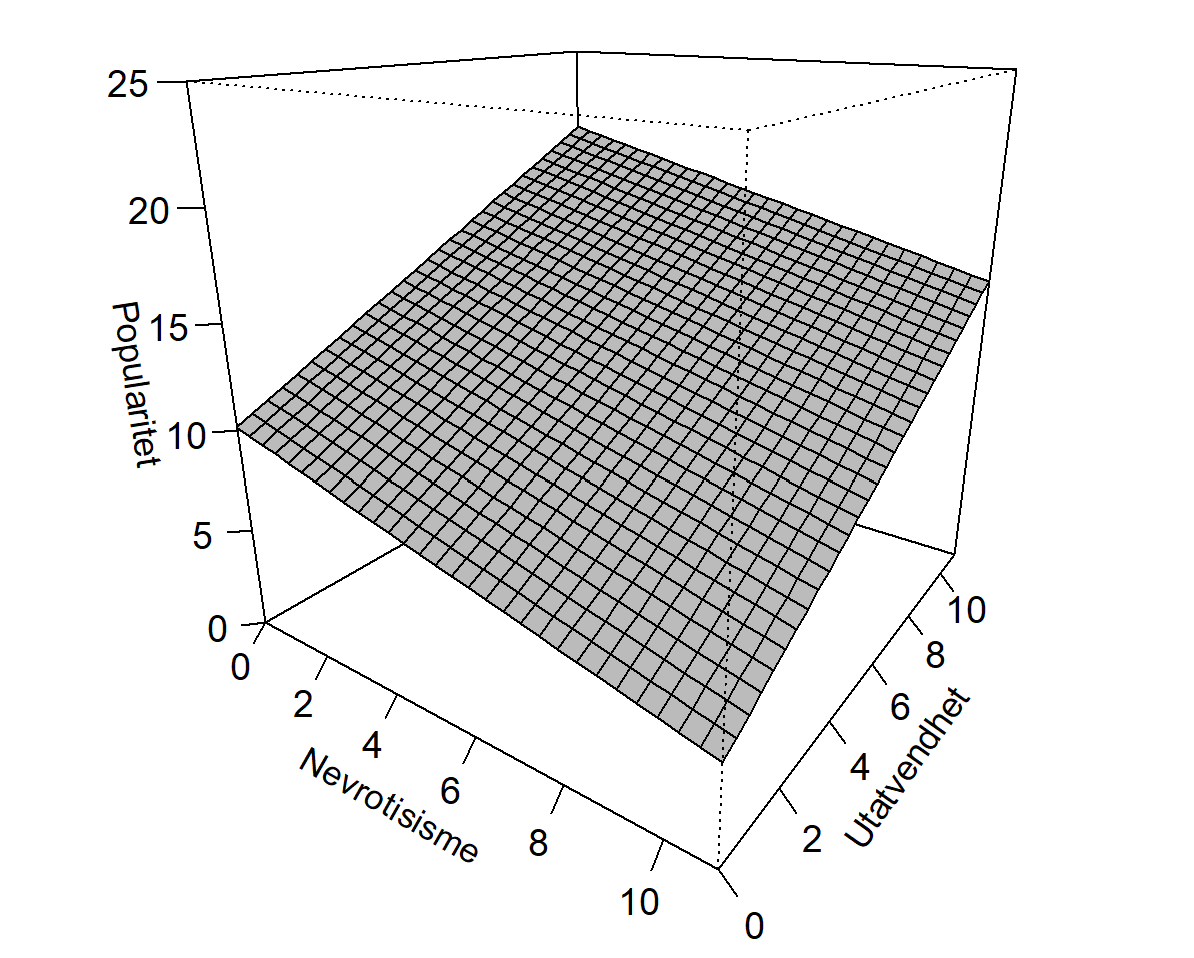


1. c. om stigningstallet er -0.5, skjæringspunktet er 10, så må den ha sunket til 5 over 10 punkter på en akse.



1. b. om det er stigningstall nærmere 0, så ser man ikke noen særlig stigning/minskning
2. En multippel regresjonsmodell har predikert y-verdi; Y^=8+3X1+5X2−4X3Y^=8+3X1+5X2−4X3 Hva forventer du skjer med YY dersom du øker X3X3 med en enhet, mens du holder X1X1 og X2X2 konstant?
   * synker med 4 enheter. Du ganger med stigningtallet, og da er den 4 enheter over.

## 3. Koeffisientene til et regresjonsplan



## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) ??? - - <2e-16 \*\*\*

## Nevrotisisme ??? - - <2e-16 \*\*\*

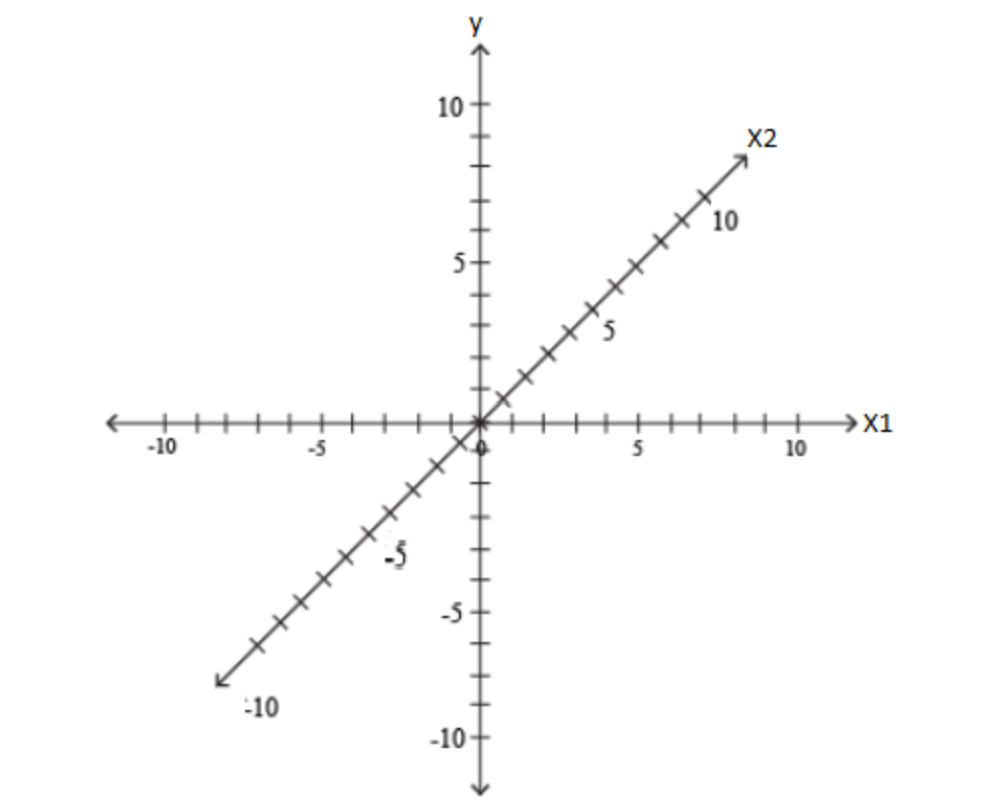
## Utadvendthet ??? - - <2e-16 \*\*\*

## ---

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value |  |
| (Intercept) | 10 |  |  |  |
| Nevrotisisme | -0,5 |  |  |  |
| Utadvendthet | 1 |  |  |  |

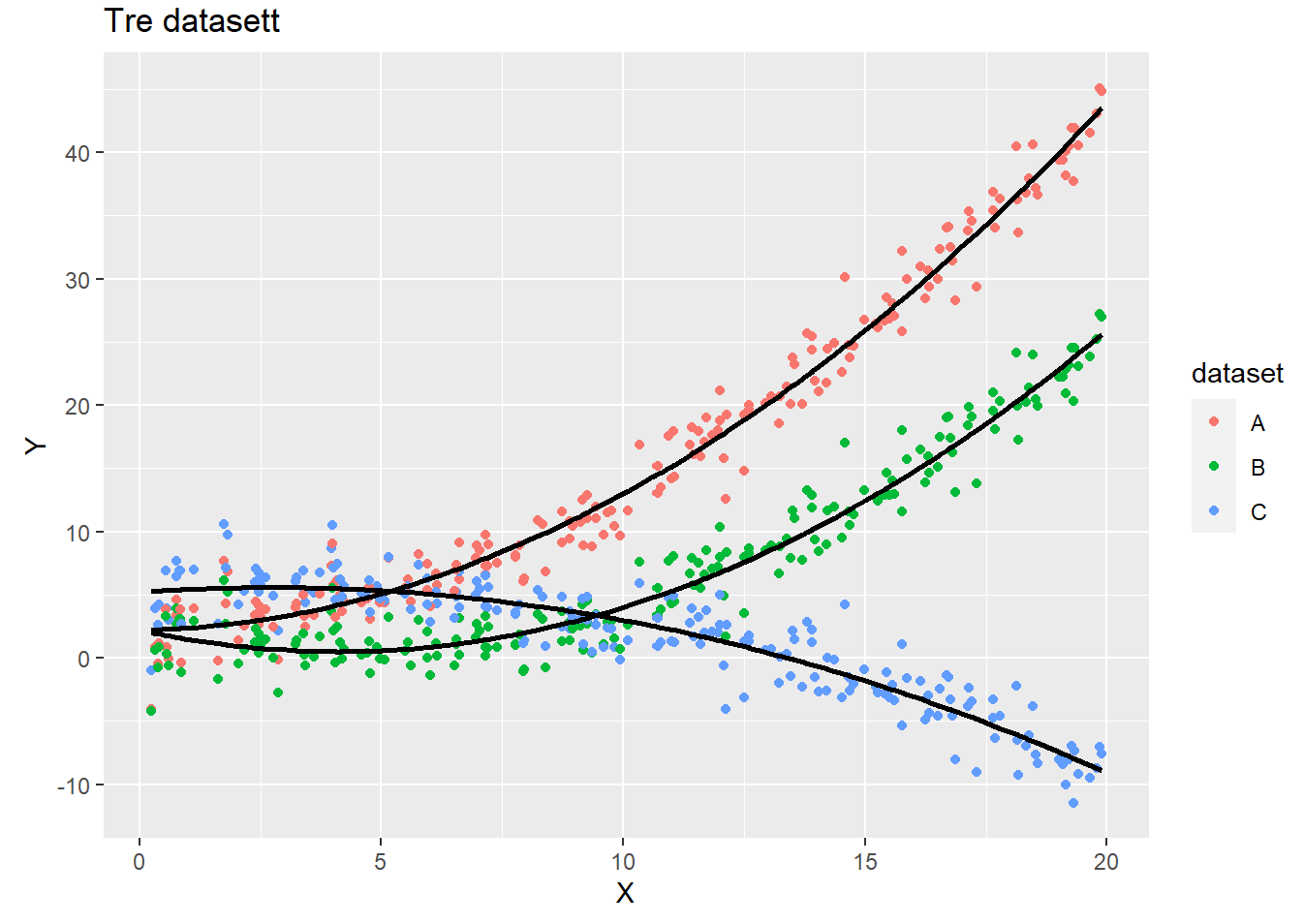
Per en enhets økning øker popularitet med én enhet

## 4. Skissere et regresjonsplan



En multippel regresjonsmodell har predikert y-verdi E(Y|X1,X2)=−1+0.5X1+1X2E(Y|X1,X2)=−1+0.5X1+1X2 Prøv å skissere regresjonsplanet til disse estimatene i koordinatsystemet over (se de geometriske figurene vist på forelesning 3. Det er ikke viktig at figuren blir nøyaktig, men tenk på hvordan dette planet er orientert i rommet). Hvordan øker Y-verdien når vi beveger oss langs X1X1 aksen kontra langs X2X2 aksen? Hva er YY i punktene (X1=0X1=0, X2=0X2=0), (X1=6X1=6, X2=0X2=0), (X1=0X1=0, X2=6X2=6), (X1=6X1=6, X2=6X2=6)?

## 5. Polynom regresjon



Under ser du et plot av tre forskjellige datasett, og forventet Y-verdi fra en kvadratisk modell av X. Hvilke av de tre datasettene A,B og C er basert på resultater fra analysene 1,2,og 3 under?

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.218608 0.403988 5.492 1.22e-07 \*\*\*

x 0.070309 0.091868 0.765 0.445

I(x^2) 0.100785 0.004421 22.795 < 2e-16 \*\*\*

* + Analyse 1: a, stiger fra start

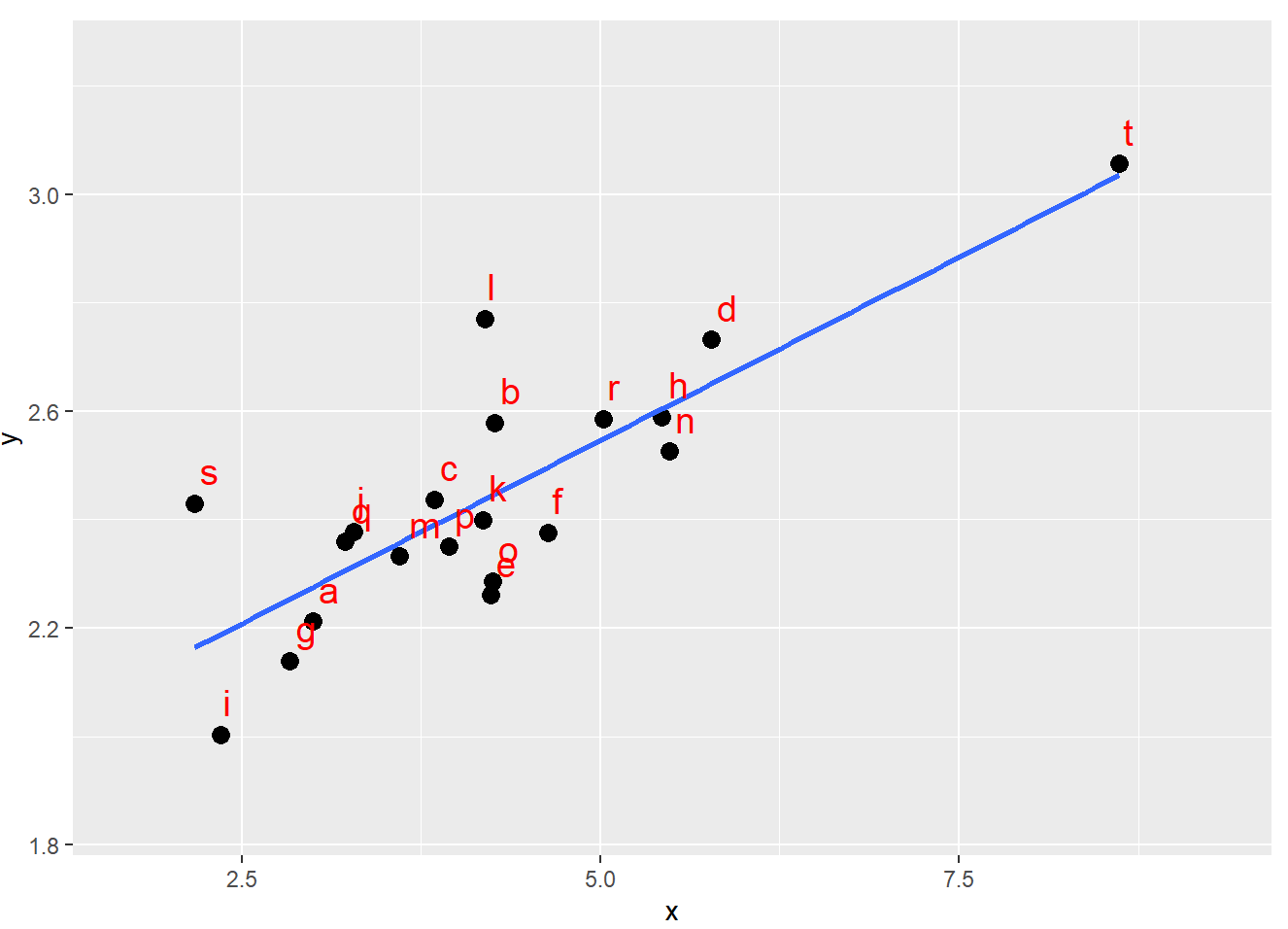
1. Coefficients:
2. Estimate Std. Error t value Pr(>|t|)
3. (Intercept) 2.218608 0.403988 5.492 1.22e-07 \*\*\*
4. x -0.829691 0.091868 -9.031 < 2e-16 \*\*\*
5. I(x^2) 0.100785 0.004421 22.795 < 2e-16 \*\*\*
   * Analyse 2: b, synker litt først før den stiger
6. Coefficients:
7. Estimate Std. Error t value Pr(>|t|)
8. (Intercept) 5.218608 0.403988 12.918 < 2e-16 \*\*\*
9. x 0.270309 0.091868 2.942 0.00365 \*\*
10. I(x^2) -0.049215 0.004421 -11.131 < 2e-16 \*\*\*
    * analyse 3: c, de to andre har likt skjæringspunkt.

Regne på de punktene som er tydelige.

## 6. Influensstatistikker

I figuren under ser du et spredningsplot der punktene er merket med bokstaver. Den blå linjen viser minste kvadraters regresjonslinje i en modell der Y er avhengig variabel, og X er uavhengig variabel. Hvilket av punktene:

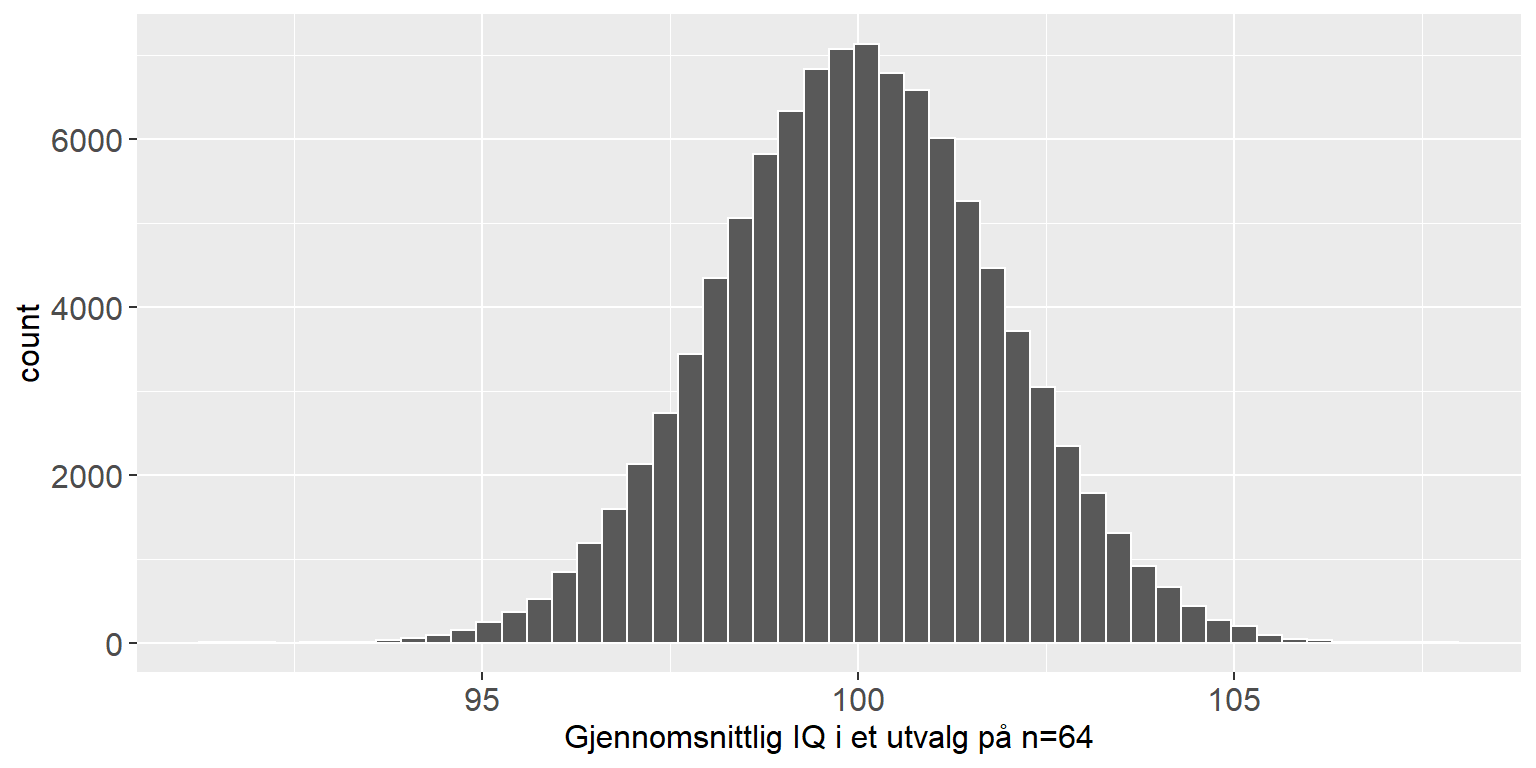
1. Hvilke punkter i plottet over tror du har størst og minst *leverage*?
   * Størst leverage: de lengst unna gjennomsnittet på x, l
   * Minst leverage: de nærmest gjennomsnittet på x, k eller p
2. Hvilke punkter tror du har størst og minst Har den største verdi av *residualene* (uavhengig av om fortegnet er positivt eller negativt, dvs. om det er over eller under regresjonslinjen)?
   * Størst verdi av residualene: de lengst unna regresjonslinja, l
   * Minst verdi av residualene: de nærmest regresjonslinja, h
3. *Cook’s distance* (Cooks avstand) til en observasjon angir hvor mye regresjonslinjen vil endre seg dersom observasjonen fjernes og analysen kjøres på nytt. Jo større verdi, jo mer vil linjen endres. Hvilke punkter tror du har størst og minst verdi for Cook’s distance.
   * s, de i starten på x-aksen vil ha større påvirkning enn de midt i for eksempel. Også de på enden for så vidt.



# Seminar #4

## 1. Samplingfordelinger

Med *kritisk verdi* mener vi den verdien som en teststatistikk må overstige for at nullhypotesen skal kunne forkastes. To viktige sannsynlighetsfordeler som brukes i inferens ved regresjon og variansanalyse er t-fordelingen og F-fordelingen. De første oppgavene er ment å gi deg litt erfaring med å bruke de relevante tabellene for å finne kritiske verdier.



I befolkningen (populasjonen) er IQ normal fordelt med gjennomsnitt 100 og standardavvik 15. Tenk deg at du trekker et utvalg på 64 personer, regner ut gjennomsnittlig IQ i utvalget. Så gjentar du prosessen på nye utvalg av samme størrelse 100.000 ganger, og lager et histogram av de verdiene for gjennomsnittet du ser i de ulike utvalgene. Et slikt histogram er gitt over.

1. Hva vil du ut ifra figuren gjette er gjennomsnittlig verdi for fordelingen? Er det rimelig?
   * IQ gjennomsnitt: 100, balansepunktet for fordelingen. Det er rimelig fordi det er sannsynlig at man får flest utvalg som er rundt gjennomsnittet.
2. Hva vil du gjette at standardavviket er? Hvorfor er standardavviket ikke det samme som det er for IQ i befolkningen (altså 15)?
   * Sd IQ sampling: ca. 2,5. standardavviket er mye lavere enn standardavviket for IQ i befolkningnen (sd befolkning = 15). Det er mye lavere fordi de ekstreme verdiene er det sjeldnere å få gjennomsnitt med.
3. Hvilket intervall (rundt gjennomsnittet) ville du basert på histogrammet ovenfor tenke at omtrent 95% av gjennomsnittene faller i?
   * 96-104, sannsynlig å få et utvalg som ligger innenfor dette intervallet
4. Vi kaller fordelingen du ser i histogrammet «samplingfordelingen til gjennomsnittet». Samplingfordelingen vil være tilnærmet normalfordelt med gjennomsnitt på 100 og standardavvik på 1.875 (standard error). [Dette ble tatt opp i PSY1010. Bakgrunnen er ikke så viktig for oss i PSY2014, men om du lurer på hvorfor standardfeilen er 1.875, kan du lese avsnitt 5.3 i Agresti]. Hvis det observerte gjennomsnittlige IQ blant 64 psykologistudenter som tar PSY2014 er 105, er det sannsynlig at PSY2014-studenter har samme IQ som den generelle befolkningen?
   * Vi må finne ut om IQ til sample ligger innenfor to standardavvik.
   * 105 (sample psy2014)-100(gjennomsnittet til populasjonen)= 5
   * Finne ut hvor mye 5 er i standardavvik: 5/1,875=2,667. psy2014s gjennomsnitt ligger 2,667 standardfeil fra populasjonens gjennomsnitt. I noen tilfeller vil dette skje ved ren tilfeldighet, men det er så usannsynlig at dette skjer ved en tilfeldighet siden vi har alpha: 0.05 og dette er en ekstrem verdi. Nullhypotesen er forkastet.
   * PS: Standardavviket i en samplingfordeling kalles standardfeil (standarderror).

## T-fordelingen

Noen teststatistikker følger en t-fordeling under H0, altså når nullhypotesen er sann vil vi se ulike verdier for statistikken i forskjellige utvalg, og det vi vil se på tvers av ulike utvalg følger en t-fordeling med et bestemt antall frihetsgrader. Denne oppgaven skal gjøre deg litt mer kjent med begrepene kritiske verdier og t-fordelinger.

1. Anta at du har en teststatistikk som under nullhypotesen følger en t-fordeling med 5 frihetsgrader. Hva er den kritiske verdien for en to-halet test med α=0.01α=0.01?
   * Se på tabell: se på df 5 og se på 0.005 siden det er en tohalet test. Da fordeler man 0.01 på to haler, og 0.01/2=0.005.
   * Kritisk verdi: 4.032
2. Hva er kritisk verdi for en to-halet test på alfa = 0,05, forutsatt at teststatistikken er t-fordelt med 10 frihetsgrader?
   * Nå ønsker vi at 0.05 skal være fordelt på halene. 0.05/2=0.025.
   * Krisitks verdi: 2.228. Om vi får en større verdi enn 2,228 eller mindre enn -2,228, så forkaster vi H0.
3. En statistisk test følger en t-fordeling med 15 frihetsgrader under H0. Du observerer verdien 2,5 for teststatistikken i dataene dine. Kan du forkaste H0?
   * Df=15 gir kritisk verdi på 2.131 med alpha=0,025. Teststatistikken 2,5 er høyere enn den kritiske verdien, så derfor kan vi beholde H0.
   * Med alpha=0.01 får man en kritisk verdi på 2,947. Teststatistikken vår er lavere enn den kritiske verdien, vi forkaster H0.
   * Forkaster H0 hvis teststatistikken er mer enn den kritiske verdien.
     + På den negative siden: < -4.032
     + På den positive siden: > 4.032
4. Legg merke til hvordan den kritiske verdien for en t-fordeling nærmer seg kritisk verdi for en standard normalfordeling (Z), ettersom antall frihetsgrader øker. Hvor mange frihetsgrader mener du (omtrent) at du trenger før de kritiske verdiene er omtrent de samme for en t-fordeling og en standard normalfordeling?
   * 1,96 på et uendelig antall frihetsgrader for t=0.025. på 30 frihetsgrader er den lik en normaldistribusjon.
   * Bare for å gjøre det litt enklere for oss når vi har normalfordeling sier vi at 95% av fordelingen ligger innenfor +-2 standardavvik, men om man skal være helt nøyaktig så er det 1,96 standardavvik. Vi bare runder til 2 fordi det er enklere å forholde seg til.
   * Frihetsgrader henger sammen med antall deltagere. df=n-p-1

## Utregning av teststatistikker i en bivariat regresjon

I oppgavesett 2 regnet vi ut minste kvadraters estimater for regresjonskoeffisientene b^0 og b^1. I denne oppgaven skal regne ut de viktigste statistikkene for inferens i bivariat regresjon.

1. Regn ut standardfeilen for b^1 og verdien til t-statistikken.
   * Et mål på hvor presis b1 er basert på de verdiene vi har fra vårt utvalg. Standardfeilen til b1 bruker vi til å regne ut test-statistikken som da er t-statistikken.
   * Regne ut standardfeil til b1: SE(b1) = S/sqrt(sigma(x-mean(x))^2
   * (SSE=summen av de kvadrerte residualene)
     + Først finne S:
       1. S = sqrt(SSE/df\_SSE) = sigma(x-x^hat)^2 / n-p-1
       2. S = sqrt(-1.0^2+(-0.5)^2+1.5^2+(-3.5)^2+(3.5)^2 / 5-1-1) = sqrt(28/3) S = 3,055
     + SE(b1) = S/sqrt(sigma(x-mean(x))^2 = 3,055/4 = 0,76
   * t = b1/SE(b1) = 1,5/0,76 = 1,96
2. Hva er null-hypotesen (H0) som vi vurderer ved bruk av t-statistikken over?
   * b=0, altså det er ingen forskjell mellom gjennomsnittene i populasjonen og utvalget.
3. Hvilken fordeling følger verdiene for t-statistikken i ulike tilfeldige utvalg under H0H0?
   * df = n-0-1= 5-1-1 =3
4. Kan du forkaste null-hypotesen gitt dataene i tabellen over?
   * Kritisk verdi =3,182 for tohalet test med alpha=0.05 (deler på to).t-verdien vår er lavere enn den kritiske verdien, derfor kan vi forkaste H0. teststatistiken vår er mindre ekstrem.
5. Finn og tolk et 95% *konfidensintervall* for b^1.
   * Lett å tolke konfidensintervall feil.
   * 95% CI: b^hat1 +- t\_df,alpha \* se(b^hat1)
   * Øvre: 1,5 + 3.812 \* 0.76 = 3.926
   * Nedre: 1,5 - 3.812 \* 0.76 = -0.926
   * Hvis det er null inne i konfidensintervallet, så kan 0 være en mulig verdi for b1, og det betyr at vi må beholde H0.
   * H0 er sann er hvis t verdien er mindre ekstrem enn den kritiske verdien.

# Seminar #5

**1. I en multippel regresjonsanalyse er en sterk korrelasjon mellom de uavhengige variablene referert til som**

1. Homoskedastisitet
2. Linearitet
3. Multikollinearitet
4. Indeterminisme
5. Justert R2

Vi vil ikke at det skal være altfor høy korrelasjon mellom de uavhengige variablene, for da vet vi ikke hvilken variabel som har påvirkning.

**2. Interaksjon er når relasjonen mellom en variabel**X1**og en variabel**YY

1. Er null inntil en tredje variable X2 er kontrollert for.
2. Ikke er statistisk signifikant
3. Er statistisk signifikant men ikke veldig sterk
4. Er ulik for ulike verdier av en tredje variabel X2.
5. Ikke kan tolkes grunnet høy korrelasjon mellom X1 og X2.

**3. Dersom du har en p-verdi på 0.02, hva betyr dette?**

1. At sannsynligheten for at nullhypotesen er riktig er 0.02
2. At sannsynligheten for at nullhypotesen er feil er 0.02
3. At sannsynligheten for få observert verdi på test-statistikken grunnet samplingfeil dersom nullhypotesen er riktig er 0.02.
4. At du ikke kan forkaste H0.
5. Ingen av disse

**4. Dersom SSE i en regresjonsmodell er 60, og SSM er 140, hva er**R2**?**

1. 0.429
2. 0.300
3. 0.700
4. 0.837

R^2= 1 – SSE/TSS

TSS=SSM+SSE

TSS=60+140=200

R^2= 1 – 60/200 = 1 – 0.3 = 0.7

**5. En regresjonsanalyse med 6 uavhengige variabler og 27 observasjoner har den kritiske verdien for t teststatistikken**

1. 27 frihetsgrader
2. 26 frihetsgrader
3. 21 frihetsgrader
4. 20 frihetsgrader

n-p-1 bestemmer formen på kurven

n: observasjoner

p: uavhengige variabler

27-6-1=20

**6. En regresjonsmodell basert på 30 observasjoner ga følgende estimater:**y=60+2.8X1+1.2X2–X3y=60+2.8X1+1.2X2–X3**. Dersom TSS var 800 og SSE var 200, hva er verdien på F-statistikken?**

1. 26.00
2. 7.69
3. 3.38
4. 0.039

t-statistikken brukes for å teste om en enkel regresjonskoeffisient er 0. t => bx =0

f-testen brukes i multippel regresjon, tester om alle variablene er 0. F => bx1 = bx2 = 0. Hvis den er signifikant, betyr det at én av disse er null. Ser på hele modellen samtidig, sjekker om alle stigningstallene er null. Om vi kan forkaste den

Ser først på F-testen først, og så går til t-test.

om én av dem er signifikant, så kan vi

F = MSM/MSE

SSM=TSS-SSE=800-200=600

F=MSM/MSE= 200/7,69=26

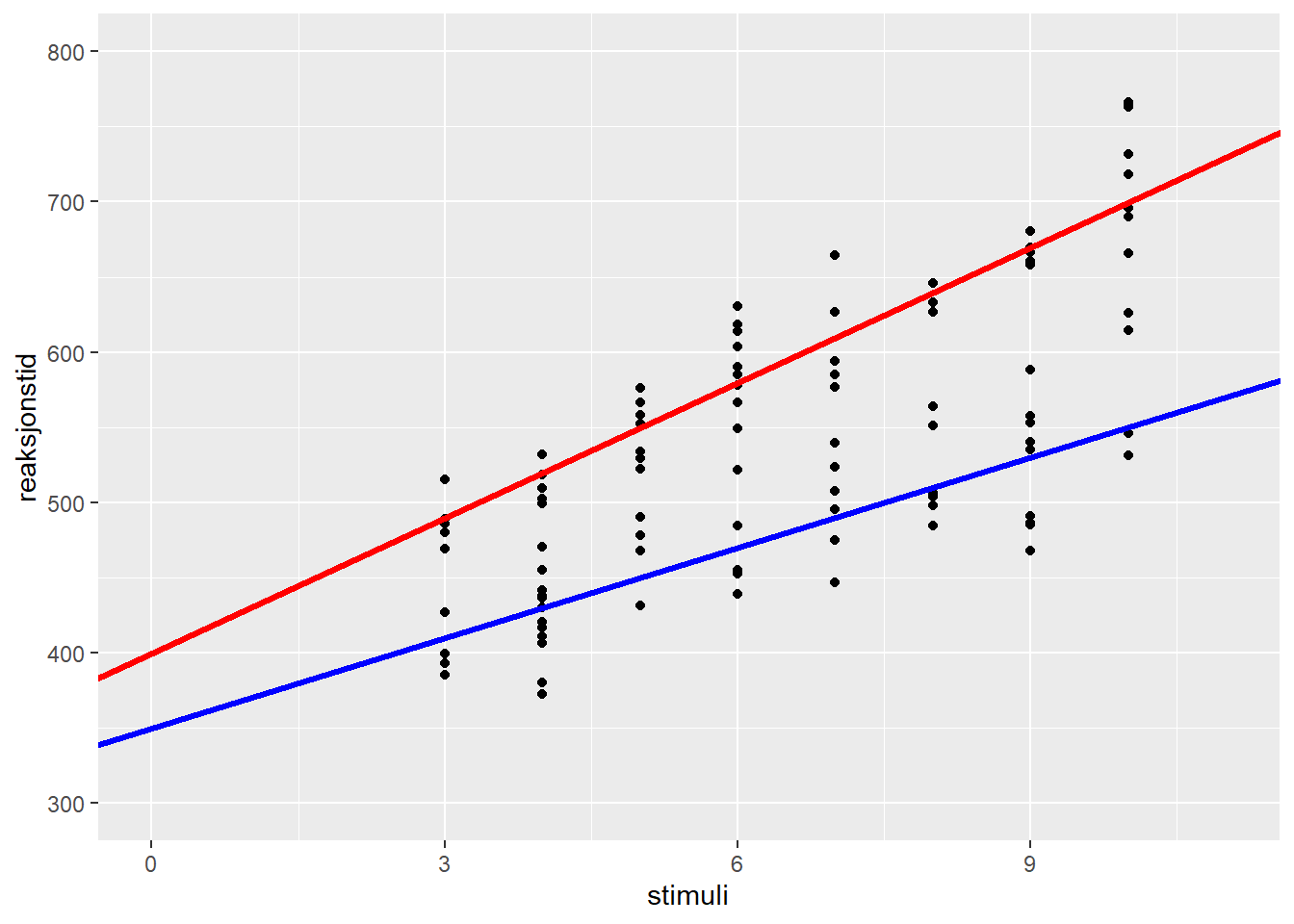
## 2. Interaksjon

1. En studie har som mål å vurdere om erfaring med dataspill henger sammen med reaksjonstiden på en visuell oppgave. Oppgaven består i at deltagere skal trykke på en knapp dersom symbolet *T* er blant et sett av visuelt liknende distraktorer. Totalt 100 personer deltar i studien.

Datasettet inneholder følgende uavhengige variabler:  
stimuli: Antall visuelle distraktorer på skjermen (fra 3 til 10)  
gamer: Kodet 1 om deltageren regelmessig spiller dataspill, og 0 ellers.

Dataene er plottet under. Den forventede reaksjonstiden for gamere er gitt i blått, og for ikke-gamere i rødt.

1. Bruk figuren til å fylle inn tallene i tabellen under som er maskert med ??? (de trenger ikke være veldig presise).



Call:

lm(formula = reaksjonstid ~ gamer \* stimuli, data = dat)

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) ??? . . .

## gamer ??? . -1.790 .

## stimuli ??? . 16.305 .

## gamer:stimuli ??? . -4.46 .

## ---

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 400 |  |  |  |
| gamer | -50 |  | -1.790 |  |
| Stimuli | 30 |  | 16.305 |  |
| gamer:stimuli | -10 |  | -4.46 |  |

RT= b0 + b1xgamer +b2xstimuli + b3Gamerxstimuli

RT ikke gamer= b0 + b2xStimuli

RT ikke gamer = 400 + 30 x stimuli

RT gamer = b0 + b1xGamer + b2xStimuli + b3GamerxStimuli

RT gamer = 400 + b1 x 1 + 30xStimuli + b3 1xStimuli

For å få riktig intercept for gamere, så må vi plusse på -50 siden den er -50 under ikke gamer.

RT gamer= 400+ (-50 x 1) + 30 xStimuli + 1 x Stimuli

RT gamer = 350 + 30 xStimuli + Stimuli

Gamer har økt fra 350 🡪 550 på y-aksen over 10 på x-aksen

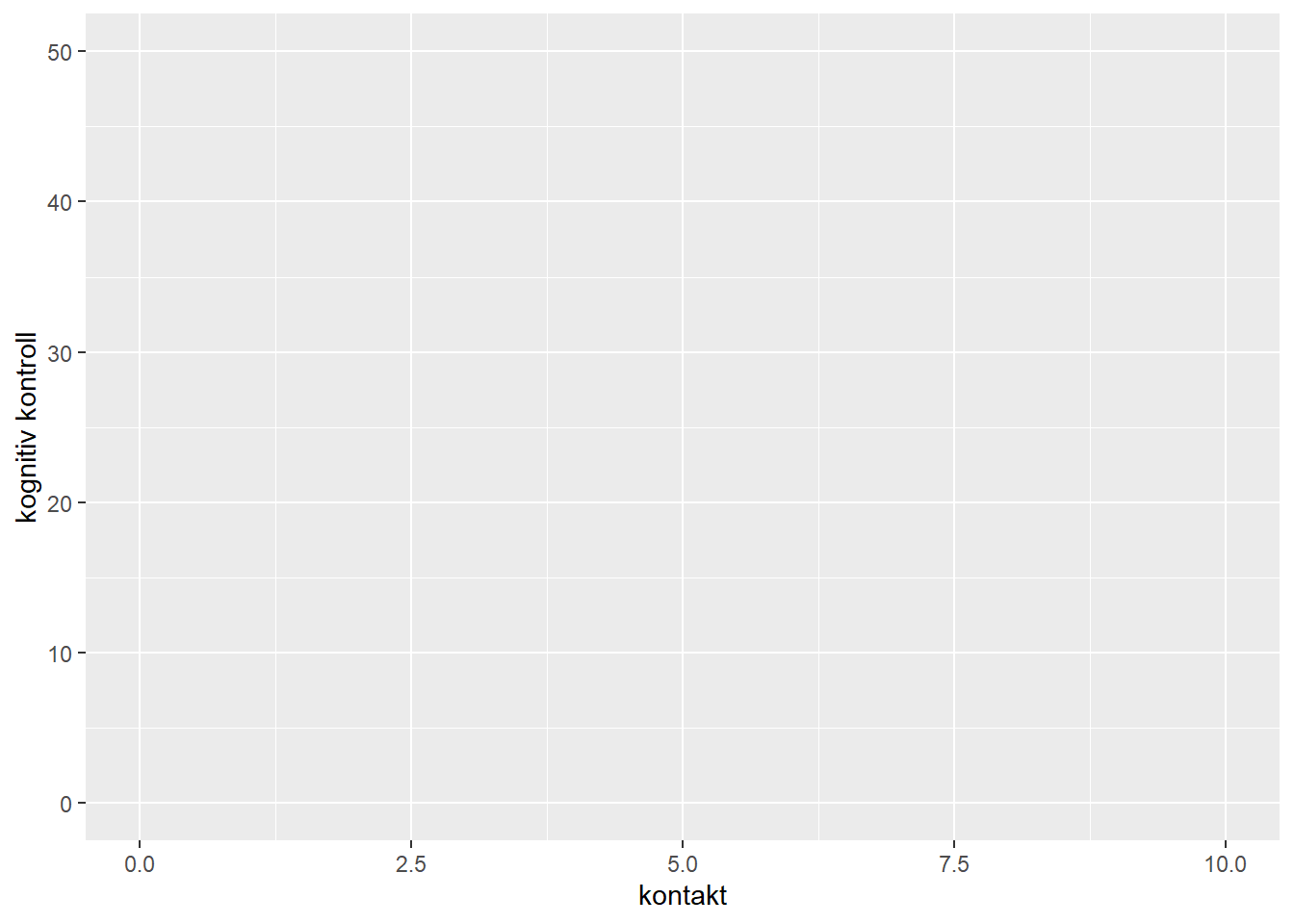
550-350=200

200/10=20

1. Hvilke hypoteser kan du teste med t-verdien til b1^ (gamer), b2^ (stimuli), og b3^ (gamer:stimuli), og hva ville du på bakgrunn disse statistikkene konkludere om forholdet mellom dataspill og reaksjonstid?
   * Vi ser at p-verdien til stimuli og interaksjon mellom gamer og stimuli er signifikante, mens p-verdien til gamer ikke er signifikant. Dermed kan de to førstnevnte forkastes, men ikke sistnevnte.
2. Koeffisientene i tabellen under kommer fra en tenkt studie av kognitiv kontroll (kontroll) hos 2 år gamle barn. kontroll er den avhengige variabelen i regresjonsanalysen.

Datasettet inneholder følgende uavhengige variabler:  
kontakt: Gjennomsnittlig timer daglig hudkontakt barnet opplever de første to ukene etter fødsel  
prematur: Premature barn er kodet 1, mens barn født til termin er kodet 0.

Bruk estimatene i tabellen under til å skissere forventet skåre for kognitiv kontroll hos premature og ikke-premature barn som funksjon av kontakt.



Call:

lm(formula = kontroll ~ kontakt \* prematur, data = dat)

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 30.0 . . .

## kontakt 0.2 . . .

## prematur -7.0 . . .

## kontakt:prematur 0.5 . . .

## ---

## 3. Indirekte effekter

En forskergruppe ser på sammenhengen mellom traume, aktivitet i hjernestammen, og hukommelse. Følgende regresjonsmodeller kjøres:

lm(hukommelse~traume)

lm(aktivitet\_hjernestamme~traume)

lm(hukommelse~aktivitet\_hjernestamme)

lm(hukommelse~traume+aktivitet\_hjernestamme)

Under finner du de estimerte koeffisientene fra de fire modellene over. På bakgrunn av disse analysene, hva ville du konkludere om sammenhengen mellom traume, aktivitet i hjernestammen og hukommelse?

---------------------------------------------------------------

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.001467 0.128672 0.011 0.991

traume -1.392355 0.126696 -10.990 <2e-16 \*\*\*

---------------------------------------------------------------

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.01248 0.10386 0.12 0.905

traume 1.10873 0.10227 10.84 <2e-16 \*\*\*

---------------------------------------------------------------

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.01527 0.07923 0.193 0.848

aktivitet\_hjernestamme -1.13723 0.05196 -21.887 <2e-16 \*\*\*

---------------------------------------------------------------

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.01388 0.07706 0.18 0.8574

traume -0.28915 0.11252 -2.57 0.0117 \*

aktivitet\_hjernestamme -0.99501 0.07494 -13.28 <2e-16 \*\*\*

---------------------------------------------------------------

Analyse 1: traume er signifikant assosiert med hukommelse

Analyse 2: traume er signifikant assosiert med aktivitet i hjernestamme

Analyse 3: aktivitet i hjernestamme er signifikant assosiert med hukommelse

Analyse 4: både traume og aktivitet i hjernestamme er signifikant assosiert med hukommelse. Når vi kontrollerer for aktivitet i hjernestamme blir traume vesentlig mindre, men er fortsatt signifikant. Her ser vi at aktivitet i hjernestamme medierer traume. Det vil si at sammenhengen mellom traume og hukommelse i stor grad er påvirket av aktivitet i hjernestammen.

Traume mediert av hjernestamme aktivitet. Indirekte effekt av traume på hukommelse.

## 4. Kategoriske uavhengige variabler

I denne oppgaven skal vi se på arbeidsrelatert stress, rapportert blant ansatte i tre ulike avdelinger ved et sykehus. Variabelen avdeling blir rekodet i to *dummyvariabler* som angitt i tabellen under:

| Avdeling | Original\_koding | Dummy1 | Dummy2 |
| --- | --- | --- | --- |
| Hjertemedisinsk | 1 | 1 | 0 |
| Alderspsykiatrisk | 2 | 0 | 0 |
| Revmatologisk | 3 | 0 | 1 |

Bruk utskriften i tabellen under til å besvare følgende:

1. Hva er forventet stressnivå i hver av avdelingene?
   * Stress^hat = b0^hat + b1 dummy1^hat + b2dummy2^hat
   * Stress\_hjerte = 40.36 + 8.573 x dummy1 + 4.355dummy2
   * De har kodet 0 på dummy 2, så det siste leddet kan kuttes
   * Stress\_hjerte = 40.26 + 8573 x 1 = 48.833
   * Stress\_alder = 40.26
   * Stress\_rev = -4.355 x 1 = 35.905
   * Man slår variablene av og på.
2. Hvilken hypotese kan testes ved og t-verdiene og F-verdien, og hva kan du her konkludere med angående stressnivået på de tre avdelingene?
   * Vi tester om H0 er sann.
   * Vi tester om dummy er likt interceptet
   * Er det forskjell i stressnivå basert på avdeling?
   * F-verdien er stor
   * Vi må beholde H0 fordi p-verdien for dummy2 er større enn alpha på 0.05
3. Hva hadde de tre koeffisientene i modellen blitt estimert til dersom du byttet verdiene som *Revmatologisk* og *Alderspsykiatrisk* avdelingde har for de to dummyvariablene?
   * Rev. og alder vil bli referansekategorier. Det vil være interceptet til stress til rev, 35.05. alderpsykiatrisk og rev. kan
   * Stressnivå i forhold hjerte vs rev. blir dummy1
   * Sammenligner med en annen referansekategori

##

## Call:

## lm(formula = stress ~ Dummy1 + Dummy2)

##

## Residuals:

## Min 1Q Median 3Q Max

## -21.2805 -4.8374 -0.1399 6.1625 23.6325

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 40.260 1.657 24.303 < 2e-16 \*\*\*

## Dummy1 8.573 2.343 3.660 0.000432 \*\*\*

## Dummy2 -4.355 2.343 -1.859 0.066435 .

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 9.073 on 87 degrees of freedom

## Multiple R-squared: 0.266, Adjusted R-squared: 0.2492

## F-statistic: 15.77 on 2 and 87 DF, p-value: 1.435e-06

5. Sammenlikning av modeller

I en studie om prediktorer for selvkontroll hos barn, testes følgende to lineære regresjonsmodeller:

m1<-lm(SELVKONTROLL~SPRÅK)

m2<-lm(SELVKONTROLL~SPRÅK+KJØNN+FORELDRES\_UTD+SØSKEN+SØVN)

1. Fyll inn verdiene maskert med *???* i utskriften under. For p-verdien kan du bare indikere om den er signifikant på et 0.05 nivå. Dere finner en tabell med F-verdier [her](https://www.stat.purdue.edu/~jtroisi/STAT350Spring2015/tables/FTable.pdf).
   * s. 341 Agresti, comparing regression models
   * formel på s. 342
   * F = MSM / MSE = 273.5 / 11.27 = 24.267
   * MSM = SSM / df\_SSM = 273.5 / 1
   * MSE = SSE / df\_SSE = 33358.4 / 298 = 11.27
   * SSE er det samme som residualenes sums of squares
   * 298 = n-1-p=300-1-1= 298
   * F(n, n-p-1)= F(1, 298)
   * Forkaste H0 fordi det er en såpass stor sammenheng.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | SS | MS | F value | Pr(>F) |
| SPRÅK | 1 | 273.5 | MSM=  SS/df=  273.5/1=  275.5 | MSM/MSE=  275.5/11.27=  24.267 | P<0.05 |
| Residuals | 298 | 3358.4 | MSE=  SS/df=  3358.4/298=  11.27 |  |  |

> anova(m1)

Analysis of Variance Table

Response: SELVKONTROLL

Df Sum Sq Mean Sq F value Pr(>F)

SPRÅK 1 273.5 ??? ??? ???

Residuals 298 3358.4 ???

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

1. Sammenlign modell 1 og modell 2 ved bruk av F-testen. Hva ville du konkludere med?

> anova(m1,m2)

Analysis of Variance Table

Model 1: SELVKONTROLL ~ SPRÅK

Model 2: SELVKONTROLL ~ SPRÅK + KJØNN + FORELDRES\_UTD + SØSKEN + SØVN

Res.Df RSS Df Sum of Sq F Pr(>F)

1 298 3358.4

2 294 2749.0 ??? ??? ??? ???

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

* + Sammenligne med en kritisk verdi for å finne ut om det er nyttig å legge til flere variabler
  + Vi har fått redusert error såpass mye
  + S. 341

# Seminar #6

## 1. Sannsynlighetsregning

Anta at du har en rettferdig terning. Bruk reglene introdusrt på forelesning 6 til å regne ut følgende utfall:

1. Hva er sannsynligheten for å få verdien 1 når du kaster terningen?
   * A(få 1)= {1}
   * S= {1,2,3,4,5,6}
   * P(A)= antall elementer i A / antall elementer I S
2. Tallene 1,3 og 5 er oddetall, mens 2,4 og 6 er partall. Hva er sannsynligheten for å få et partall når du kaster terningen? Hvorfor kan du bruke den additive regelen for å finne denne sannsynligheten?
   * P(2 eller 4 eller 6)= 1/6 + 1/6 + 1/6 =3/6 = 1/2
3. Anta at du har to terninger, en rød og en blå. Først kaster du den røde, og så kaster du den blå. Hva er utfallsrommet (hvilke utfall er mulig)? Hva er sannsynligheten for at du får tallet 1 på den røde terningen og et partall på den blå terningen? Hvorfor kan du bruke den multiplikative regelen for å finne denne sannsynligheten?
   * S={11,12,13,15,16,21,22,23,24,25,26, 31,32,33,34,35,36,41,42,43,44,45,46,51,52,53,54,55,56,61,62,63,64,65,66}
   * P(A og B)= P(A) x P(B)
   * P(1 på det først kastet)= 6/36 = 1/6
   * P(partall andre kast)=18/36 = ½
   * P(1, partall)= 1/6 x ½ = 1/12

## 2. Kji-kvadrat analyser

| By | Audi | Toyota | Volvo | Sum |
| --- | --- | --- | --- | --- |
| Oslo | 20 | 50 | ??? 10 | 80 |
| Bergen | 10 | 30 | ??? 20 | 60 |
| Trondheim | ??? 40 | ??? 20 | ??? 30 | 90 |
| Sum | 70 | 100 | 60 | ??? 230 |

1. Anta at du er interessert i fordelingen av bilmerker i ulike norske byer. Du samler 80 observasjoner fra Oslo, 60 fra Bergen og 90 fra Trondheim. Totalt er det 70 Audier, 100 Toyotaer og 60 Volvoer i utvalget. Vis at tabellen har fire frihetsgrader, ved å sett inn de manglende verdiene i tabellen. (Siden de manglende verdiene følger av de øvrige tallene i tabellen, er de ikke «frie»).
   * df = (r-1)(c-1)= (3-1)(3-1)= 2 x 2 = 4

| Status | Veldig\_kreativ | Middels\_kreativ | Lite\_kreativ | Sum |
| --- | --- | --- | --- | --- |
| Ikke-schizofren | 99 | 222 | 154 | 475 |
| Schizofren | 19 | 70 | 36 | 125 |
| Sum | 118 | 292 | 190 | 600 |

1. Schizofreni er en mental lidelse der mange som er rammet opplever hallusinasjoner og vrangforestillinger. Det tidvis bisarre innholdet i disse symptomene har gjort at man lenge har spekulert i at det er en sammenheng mellom kreativitet og schizofreni. Under finner du en krysstabell fra en tenkt studie av 600 schizofrene og ikke-schizofrene individer.
2. Formuler nullhypotesen du ønsker å teste.
   * H0 = schizofrenistatus og kreativitet er uavhengige variabler
3. Beregn antallet personer du ville forvente å se i hver celle i tabellen dersom nullhypotesen er sann. (For eksempel, det er 99 veldig kreative ikke-schizofrene individer. Hvor mange ville du forvente å se dersom H0 var sann?)
   * i: rad
   * j: kolonne
   * Eij = Ri x Cj / n , gjøre dette for hver celle
   * 475 x 118 / 600 = 93,42
   * 475 x 292 / 600 = 24,583
   * 475 x 190 / 600 =
   * 125 x 118 /600 =
   * 125 x 292 / 600=
   * 125 x 190 /600 =
   * Om det ikke er noen sammenheng, om nullhypotesen er sann, så vil de faktiske verdiene matche de forventede verdiene.
   * O: observert verdi
   * E: estimert verdi
4. Regn ut Kji-kvadrat statistikken og finn den kritiske verdien for null-hypotesen i tabellen under. Hva ville du konkludert med om sammenhengen mellom schizofreni og kreativitet?
   * Kji-kvadrat starter i 0, så man ser på 0.05-nivå fordi det ikke vil gi mening å finne verdier på minus-sida.
   * Den kritiske verdien er 5,99 for df=2 og 0.05-nivå og derfor kan vi ikke forkaste H0.
5. Finn de standardiserte og ustandardiserte residualene for cellen *Veldig\_kreativ* og *Ikke-schizofren*. Hvordan tolker du disse?
   * Standardisert residual:

# Seminar #7

## Enveis variansanalyse av voldelige dataspill og aggresjon

Man ønsket å undersøke effekten av spilling av voldelige dataspill på barns aggresjonsnivå. Aggresjon ble observert og skåret på en kvantitativ skala etter tre ulike intervensjoner. Nedenfor finnes resultatene for 9 barn fordelt på tre grupper:

gruppe 1 spilte ikke dataspill  
gruppe 2 spilte et voldelig dataspill  
gruppe 3 spilte et problemløsningsorientert dataspill.

| Gruppe | Aggresjon |
| --- | --- |
| 1 | 6 |
| 1 | 2 |
| 1 | 1 |
| 2 | 8 |
| 2 | 6 |
| 2 | 7 |
| 3 | 3 |
| 3 | 2 |
| 3 | 1 |

1. Gjennomfør alle beregninger nødvendige for en enveis variansanalyse og fyll ut tallene som er merket med “???” i tabellen nedenfor. For et av “p”, angi om verdien er over eller under 0.05. Dere kan bruke uttrykkene for “Within-groups” og “between-groups” estimatene som dere finner i avsnitt 12.3 i Agresti.

| kilde | SS | Df | MSS | F | p |
| --- | --- | --- | --- | --- | --- |
| Between | ??? | ??? | ??? | ??? | ??? |
| Within | ??? | ??? | ??? |  |  |
| Total | ??? | ??? | ??? |  |  |

* + SS: sums of squares
  + MSS: mean squared between og mean squared within??
  + Enveis variansanalyse: da ser vi bare på mellom gruppene
  + Toveis variansanalyse: ser på innad i gruppa også. Da får man enda en faktor
  + Er det større aggresjon mellom grupper enn innad i grupper?
  + Høyere F-verdi vil gi mer sannsynlighet for at den er over en kritisk verdi, og da kan man si at det er usannsynlig å få denne statistikken om det ikke er noen forskjell mellom gruppene
  + F= variasjon mellom/variasjon innad
  + De forskjellene vi kan forklare er forklart varians, og det er forskjellene mellom grupper. Vi tenker at det er på grunn av as de er i forskjellige grupper som gjør dem forskjellige
  + Den variansen innad, kan vi ikke forklare. Feilvarians. Residual error
  + Anova: minst ett av gjennomsnittene er annerledes. Vi vet ikke hvilket, så da må vi gjøre flere tester.
  + Nullhypotese: ingen forskjell mellom gruppene
  + fremgangsmåte
    - starte med å regne ut SS\_between (variasjonen imellom)
      1. finne ut av gruppegjennomsnittet
      2. y1: 3, y2: 7, y3: 2. totalt aggresjonsgjennomsnitt: 4
      3. for hvert datapunkt tar man gjennomsnittet til gruppa og trekker fra totalgjennomsnittet. SS\_between= 42
      4. SS\_total=60
    - regne sums of squares within. Hvor forskjellig er individet forskjellig fra gruppegjennomsnittet sitt? For hver deltager trukket fra gruppegjennomsnittet og kvadrerer, og så summerer alle individer.
    - Ss\_within=18
    - For å gjør F-testen, så bruker man mean sums of square. Man må dele på frihetsgradene. Vi kan ikke bruke SS for å finne F, vi må bruke MSS.
    - Regn ut MSS
      1. MSS\_between = SS\_between/df\_between og MSS\_within=SS\_within/df\_within
      2. Df\_between=grupper-1 = 3-1 = 2
      3. Df\_within=n-grupper = 9-3= 6
      4. Vi har 8 frihetsgrader totalt
      5. MSS\_between= SS\_between/df\_between= 42/2 = 21
      6. MSS\_within= SS\_within/df\_within= 18/6 = 3
    - Gjøre F-test: F= 21/3 = 7
    - Hvor stor sannsynlighet er det å få f-verdien når nullhypotesen er sann? Finne en verdi for å si at det er såpass usannsynlig for å få 7 om det ikke er noen forskjell mellom gruppene. Vi må sjekke tabellen med kritiske verdier for F(2,6). alpha: 0.05. enhalet test.
      1. når man skal se på en f-verditabell, så er det to forskjellige frihetsgrader. Numerator for between og denominator for within.
      2. kritisk verdi for F(2,6)=5,14. forkaste nullhypotesen? Ja, fordi man fikk en verdi på 7 som er mer ekstrem enn 5,14. teststatistikken er mer ekstrem enn den kritiske verdien.
      3. vi vet ingen ting om hvilke som er forskjellige fra hverandre

1. Hvilken hypotese kan vurderes gjennom analysen i punkt 1, og hva vil dere konkludere med angående aggresjon hos spillere?
2. Bruk funksjonen aov i R til å gjennomføre variansanslysen, og sammenlikn tallene du fikk i oppgave 1. Dere kan lese inn dataene med følgende kode:

Gruppe<-factor(c(1,1,1,2,2,2,3,3,3))

Aggresjon<-c(6,2,1,8,6,7,3,2,1)

* + Det er gruppe som er between og residuals som er within

1. Dersom dere har konkludert med at denne kan forkastes, så kan det være interessant å sammenligne hvert gruppegjennomsnitt med hvert av de andre – såkalte *parvise sammenligninger* - eller *kontraster*. Hvor mange slike mulige sammenligninger kan vi gjøre her? Hvor mange kunne vi gjort dersom vi hadde 6 eksperimentgrupper?
   * Tre parvise sammenligninger
   * Jo flere grupper vi har, jo flere parvise sammenligner blir det
   * 6 grupper: 15 sammenligninger. (6\*6 - 6)/2=15

## Oppgaver fra Agresti

1. Gjør regneoppgave 12.5 og 12.7, og flersvarsoppgavene 12.39 og 12.40 i Agresti.
   * 0hypotese: testes med anova
   * Forskningshypotese: antall venner du har bestemmes av hvor ofte du går på bar
   * Antgelser, kriterier som må være oppfylt for å gjøre regresjonsanalyser og anova (s.370).
     + Normalfordelt
     + Standardavvik må være samme for hver gruppe
   * Referansegruppe: de som aldri går til bar. De skal bli sammenlignet med de som går veldig ofte til bar og de som går noen ganger på bar.
   * Dummykoding:

|  |  |  |
| --- | --- | --- |
|  | D1 | D2 |
| V0 | 1 | 0 |
| 0 | 0 | 1 |

* + Utfallsvariabel: venner
  + Intercept: verdien for referansegruppen. Gjennomsnittlig antall venner for de som aldri går på bar
  + Dummy1: forskjellen mellom antall venner for de som har 1 på dummy 1 og referansegruppen
  + Venner^hat=b0+X\*d1+X\*d2
  + V^hat= 6,2+5,9\*d1+0,2\*d2

12.7

* Hvis gruppe 3 hadde hatt gjennomsnittsskåre 9 og ikke 1 på quizen, så ville standardavvikene vært det samme, men gjennomsnittene i gruppene hadde vært 6,7,8 istedenfor 6,3,8.

|  |  |  |
| --- | --- | --- |
| A | B | C |
| 6 | 3 | 8 |
| 6 | 7 | 8 |

* + Er det større forskjeller mellom enn innad? For hvis det er det, så vil F-verdien blir høyere som gjør at vi oftere forkaster H0.
  + Mellomgruppevariasjonen ville blitt mindre med 7 enn 3, og F ville blitt mindre.
  + Fordi gruppegjennomsnittene, variasjonen mellom gruppene har blitt mindre

|  |  |  |
| --- | --- | --- |
| A | B | C |
| 2 | 2.8 | 2.6 |
| 1 | 1.8 | 1.6 |

* + Standardavvik sier noe om hvor mye variasjon det er innad i gruppene.

|  |  |  |
| --- | --- | --- |
| A | B | C |
| 3 | 2 | 3 |
| 30 | 20 | 30 |

* + Anta at vi har samme gjennomsnitt og standardavvik, men antall deltagere er 30 istedenfor ... hvis vi har økt antall deltagere fra 3,2 og 3, vil f-statistikken bli større, den samme, eller mindre?
  + Når vi har flere deltagere blir det enklere å fange opp reelle forskjeller. Vi får mer stryke (power) i testen vår.
  + Hvor mange deltagere påvirker frihetsgradene. Når vi har større frihetsgrader, og siden den er avhengig av antall deltagere, så vil den bli mindre, og vil få en større F-statistikk. Med flere deltagere vil det være enklere å oppdage de forskjellene som er tilstede.
  + Større utvalg gjør forskjeller mer synlig.
  + Når F-statistikken blir større, så blir p-verdien mindre. De henger sammen. Fordi det er en fordeling, så jo høyere den verdien er, jo mindre sannsynlig er det å få den. Fordi f-distribusjonen ser ut som ei sklie som blir smalere mot høyre.

# Seminar #8

Ved å legge til en til faktor, så får vi faktoriell analyse

Istedenfor én faktor, så skal vi ta med én til, så det blir en toveis anova.

Effekten av fargen på smertestillende tabletter  
Å senke hånden ned i et kar med isvann oppleves av de fleste som svært smertefullt, og fordi det ikke er forbundet med vevsskade er dette en vanlig måte å indusere smerte i eksperimentelle studier.

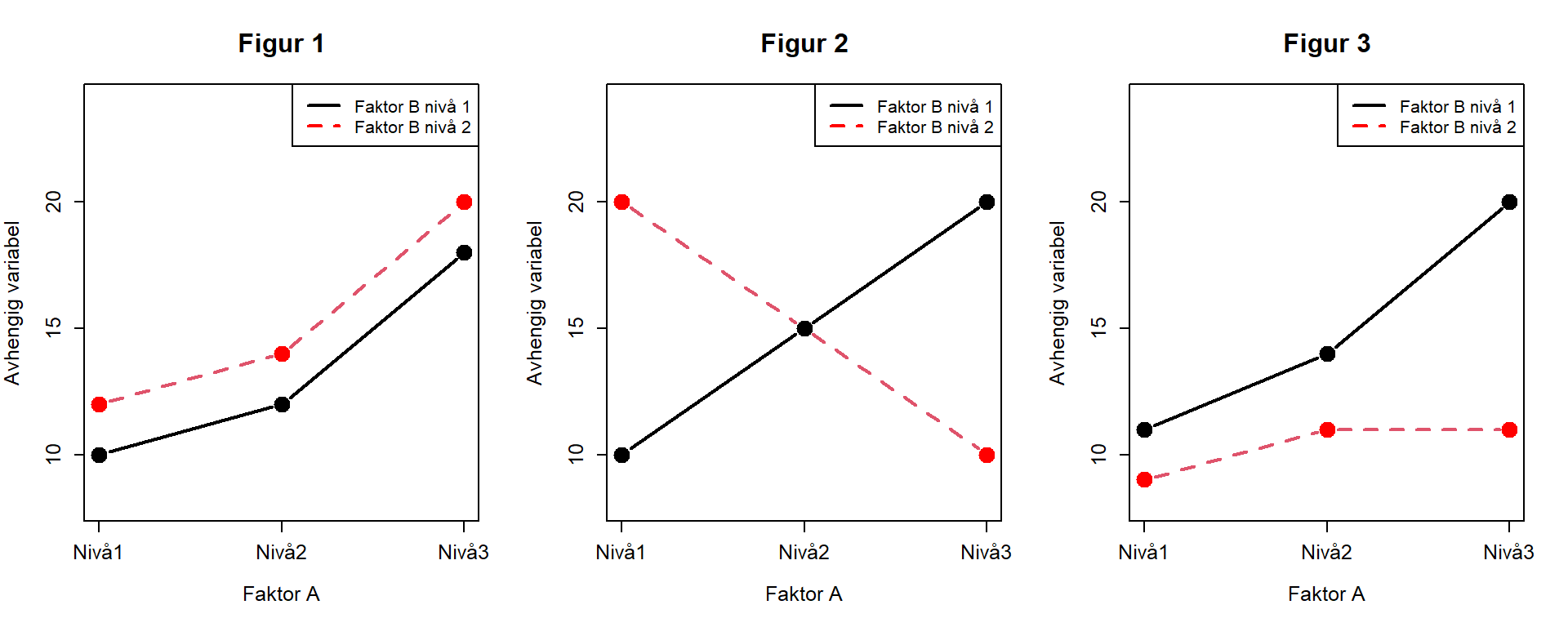
I denne studien tar deltagerne en smertestillende tablett 10 min før de senker hånden ned i isvannet. Denne tabletten kommer i en av tre farger; Rød, gul eller grønn. Den avhengige variabelen i denne oppgaven er grad av smerte rapportert av deltagere 30 sekunder etter at de senket hånden ned i vannet.

1. Først ble det gjennomført en enveis variansanalyse av farge av tablettene og smerte. Fyll ut de sladdede verdiene i tabellen under.

| kilde | SS | df | MSS | F | p |
| --- | --- | --- | --- | --- | --- |
| Between | ??? 238,576 | ??? 2 | ??? 119,288 | ??? 2,864 | ??? 0,06 |
| Within | 4873.643 | ??? 117 | 41.655 |  |  |
| Total | 5112.219 | 119 |  |  |  |

* + SS\_total = SS\_between + SS\_within
  + SS\_between = SS\_total – SS\_within = 5112,219 – 4873,645 = 238,576
  + F = variansen mellom / variasjon innad
  + For å komme fram til MSS, så må man finne ut av df
    - Df\_between = antall grupper – 1. 3-1 = 2
    - Df\_within = antall deltagere – antall gru pper. 120 – 3 = 117
    - Df\_total = antall deltagere – 1 N – 1 = 119
  + MSS\_between = 238,576 / 2 = 119,288
  + F = MSS\_between / MSS\_within = 119,288 /41,655 = 2,864
  + F(df1, df2
  + Numerator (teller): frihetsgradene between, denominator (nevner): frihetsgradene til within.
  + Hvis verdi er høyere enn cut-off, da er det usannsynlig at den er sannsynlig å få
    - Finner kritisk verdi ved hjelp av R: 3,07
    - Alpha=0.05
    - F-statistikken er mindre ekstrem enn den kritiske verdien, den ligger under den kritiske verdien. Det er mer sannsynlig å få den dersom H0 er sann.
    - Fargen på pillene ser ikke ut til å påvirke smerte. Vi må beholde H0.
    - P > 0.05, **ikke signifikant**. F<kritisk verdi (cut-off)
    - Noen ganger vil det være funn som ser ut som om den er signifikant, men det er den altså ikke.
    - Alpha sier noe om hvor sikre vi er på resultatet.

1. I figuren under finner du tre linjediagrammer som viser eksempler på ulike scenarioer som kan komme til uttrykk i 2-faktorielle eksperimenter (faktor A og faktor B). Sammenlikn de tre figurene, og beskriv hvilke tanker du gjør deg om de to faktorene A og B.



* + Tofaktoriell: to faktorer/variabler
    - Faktor A: farge-gruppa, rød, gul eller grønn tablett (nivå, 1,2,3)
    - Faktor B nivå 1: paracet
    - Faktor B nivå 2: placebo
  + Figur 1
    - Ser ut som om uavhengig at det er forskjeller i fargene i rapportert smerte.
      1. det ser ut som farge spiller en rolle i rapportert smerte, det er en hovedeffekt
      2. hovedeffekt av type effekt hvor de som får placeobo rapporterer høyere smerte enn de som har fått paracet
  + figur 2
    - interaksjon mellom de to faktorene: faktor a farge og faktor b type tablett.
    - Effekten av farge avhenger om du har fått placebo eller paracet.
    - Hovedeffekten av faktor A forsvinner i den interaksjonen om vi ikke tar høyde for type tablett.
    - Interaksjonseffekt mellom farge og type tablett.
  + Figur 3
    - Ser ut at det kanskje kan være en interaksjon her også.
    - Det ser ut som effekten av grønn tablett har større effekt
    - Hovedeffekt av tablett, en effekt av farge og kanskje en interaksjon (effekten av den ene faktoren er avhengig av nivået på den andre faktoren).
    - Interaksjon: Hvor effekten av farge avhenger av hvilken tablett man har fått.

1. Under finner du resultatet fra en toveis-variansanalyse med med faktorene FARGE og GRUPPE, gjennomført i R. Tolk resultatene, og sammenlikn med resultatene enveis-variansanalysen.
2. > aov.res=aov(value~farge\*gruppe, data=dat)
3. > summary(aov.res)
4. Df Sum Sq Mean Sq F value Pr(>F)
5. farge 2 239 119.3 3.651 0.02905 \*
6. gruppe 1 804 803.9 24.606 2.48e-06 \*\*\*
7. farge:gruppe 2 345 172.5 5.280 0.00641 \*\*
8. Residuals 114 3725 32.7
9. ---
10. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
    * Frihetsgrader for farge:gruppe, da tar man og ganger df for farge med df for gruppe. 2\*1=2
    * Sums of squares residual har blitt redusert fra oppg. 1. Kan forklares at vi har lagt inn tablett. Vi kan forklare mer av variasjonen pga. vi la til denne faktoren.
    * Det er mindre sannsynlighet å få f-verdien til farge dersom H0 er sann. I den første analysen fant vi at den ikke var signifikant, men etter å ha lagt til en faktor, så er den signifikant. MS\_within har blitt mindre, og F blir høyere, mer kritisk.
    * Sterk hovedeffekt av om man har fått placebo eller paracet.
    * Interaksjonsledd er også signifikant. Effekten av farge ser ut til å avhenge om man har fått paracet eller placebo.

## Oppgaver fra Agresti

1. Gjør regneoppgave 12.19, 12.22, og flersvarsoppgavene 12.42 i Agresti.

12.16

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SS | df | Mean square | F | P |
| Gender | 100 | Gruppe -1  2 – 1 = 1 | SS/df  100/1=100 | F=MSS\_between/MSS\_within  100/18=5,6 | 0.019 |
| Marital status | 200 | Gruppe -1  3-1 = 2 | 200/2=100 | 100/18=5,6 | 0.004 |
| Interaction | 100 | Gender\_df\*marital status\_df =  2\*1 = 2 | SS/df  100/2=50 | 50/6=2,8 | 0.06 |
| Residual (Error) | 4000-100-200-100 = 3600 | Antall deltagere – antall grupper =  206-6= 200 | SS/df  3600/200=18 |  |  |
| Total | 4000 | 205 |  |  |  |

* + Sidenote: Ved 30 deltagere, så begynner det å nærme seg kritisk verdi som er 2, og da kan man begynne å se på z istedenfor t-verdi.