

Oppgave 1: Prediktorer for verbal hukommelse (55%)

I denne oppgaven skal vi vurdere ulike prediktorer for verbal hukommelse eldre. Skåren på en standardisert test av verbal hukommelse er gitt i variabelen VERBAL, og høye verdier betyr at deltakeren husket mange av test ordene. VERBAL vil være den avhengige variabelen i alle analysene.

Følgende uavhengige variabler vil bli inkludert i regresjonsmodellene:

SØVN.C: Gjennomsnittlig antall minutter deltageren sover per natt. Variabelen SØVN.C er "sentrert", dvs, gjennomsnittlig søvnlengde i utvalget er trukket fra målet på søvn.

- ALDER: Alder målt i år.
- ABSTRAKT: Abstrakte substantiver referer til de vi ikke kan oppfatte direkte gjennom en av sansene våre. Variabelen er kodet 1 dersom deltageren fikk i oppgave å huske en liste abstrakte substantiver (for eksempel "RETTFERDIGHET" eller "HÅP"), og 0 dersom deltageren har lest en liste med konkrete substantiver (for eksempel "BORD" eller "HYL").
- SØVN.CxABSTRAKT: Resultatet av multiplikasjonen mellom variablene SØVN.C og ABSTRAKT (SØVN.C * ABSTRAKT).
- FYSISK: Gjennomsnittlig antall minutter i uken med fysisk aktivitet.
- SOSIAL: Gjennomsnittlig antall timer i uken med sosial aktivitet.
- HODEOMKRETS: Omkretsen på deltagerens hode, målt i cm.

a) I modell 1 er SØVN den eneste uavhengige variabelen.

1. Kovariansen mellom SØVN.C og VERBAL er 17.06, og standardavviket til SØVN.C-variabelen er 13.27. Vis at stigningsgraden til regresjonslinjen i modell 1 er lik 0.097.

$$\hat{b}_1 = \frac{s_{XY}}{s_X^2} = \frac{17.06}{13.27^2} = 0.097$$

- Hva er forventet skåre på VERBAL for en person som sover 100 minutter mer enn gjennomsnittet?

$$\widehat{VERBAL}_i = \hat{b}_0 + \hat{b}_1 X_i = 67.81278 + 0.09694 * 100 = 77.50678$$

Relevant bakgrunnsstoff å diskutere:

- I og med at variabelen er sentrert, blir \hat{b}_0 et direkte estimat på forventet verbal hukommelse hos en person som sover gjennomsnittlig antall minutter.

b) Modell 2 inneholder følgende uavhengige variabler; SØVN.C, ABSTRAKT og SØVN.CxABSTRAKT.

1. Hvordan vil du oppsummere forholdet mellom verbal hukommelse, søvn og abstrakte ord basert på modell 2?

Det viktigste å være oppmerksom på og kommentere i besvarelsen, er at vi her modellerer en interaksjon mellom abstrakte/konkrete ord og søvn. I og med at ABSTRAKT tar verdiene 0 eller 1, så kan vi eventuelt formulere svaret som to forskjellige regresjonslinjer, en for hver av gruppene i ABSTRAKT.

$$\widehat{VERBAL}_{konkret} = \hat{b}_0 + \hat{b}_1 * SØVN.C_i + \hat{b}_2 * ABSTRAKT_i + \hat{b}_3 * ABSTRAKT_i * SØVN.C_i$$

$$\widehat{VERBAL}_{konkret} = 75.06080 + 0.10176 * SØVN.C_i + (-14.57721) * 0 + 0.08406 * 0 * SØVN.C_i$$

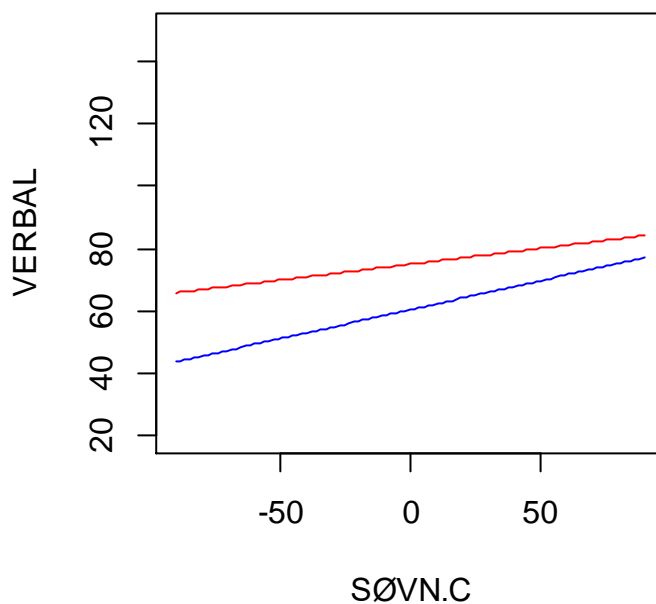
$$\widehat{VERBAL}_{konkret} = 75.06080 + 0.10176 * SØVN.C_i$$

$$\widehat{VERBAL}_{abstrakt} = 75.06080 + 0.10176 * SØVN.C_i + (-14.57721) * 1 + 0.08406 * 1 * SØVN.C_i$$

$$\widehat{VERBAL}_{abstrakt} = (75.06080 - 14.57721) + (0.10176 + 0.08406) * SØVN.C_i$$

Oppsummert med ord; Ved gjennomsnittlig søvnlengde skårer folk i snitt 75.06 på hukommelsen av konkrete ord. Ved gjennomsnittlig mengde søvn skårer gruppen som har i oppgave å huske abstrakte ord 14.577 poeng lavere enn de som husker konkrete ord.

Søvn utover gjennomsnittlig lengde er assosiert med en økt hukommelse, 0.10176 ord per min. I gruppen som husker konkrete ord er forholdet mellom hukommelse og søvn enda sterkere (ca 80% høyere, $0.10176 + 0.08406 = 0.18582$ per min.) enn for de som husker abstrakte ord. Under er dette plottet i intervallet +-1.5 timer fra gjennomsnittet. Ifølge modellen er



- Besvarelsen bør ta opp signifikansen til parameterne (alle er signifikante $p < 0.05$).
- Forklart varians (R^2) bør nevnes, og her er den 0.7452. 74.52% av variansen i verbal hukommelse kan tilskrives de tre uavhengige variablene. Mesteparten trolig til ABSTRAKT, i og med at søvn alene bare forklarer litt over 2% av variansen i modell 1.
- Informasjon fra korrelasjonstabellen burde også ta med.

c) I modell 3 blir ALDER lagt til modellen.

1. Hvordan forandrer det å legge til den uavhengige variabelen ALDER hvordan vi forstår forholdet mellom verbal hukommelse, søvn og abstrakte ord?

Kjernepoeng som bør komme frem:

Vi ser at når ALDER legges til modellen synker effekten av SØVN.C vesentlig (nesten 50%).

Ut ifra korrelasjonsmatrisen ser vi at ALDER er moderat negativt korrelert med SØVN.C, og eldre mennesker sover mindre enn yngre mennesker. SØVN er fortsatt signifikant ($p < 0.001$), men p-verdien er nødvendigvis høyere enn i modell 2.

Det er viktig at kandidaten her trekker inn konseptet «konfunderende variabel», og definerer dette. En konfunderende variabel (confounder) er en bakenforliggende variabel som korrelerer både med eksponeringen av interesse (SØVN.C) og den avhengige variabelen (VERBAL). Dette kan føre til at den observerte bivarierte korrelasjonen er «spuriøs», i den forstand at den ikke er uttrykk for en kausal relasjon mellom SØVN.C og VERBAL.

Dette kan med fordel illustreres med en figur.

Det bør nevnes at forklart varians i modell 3 er 76.57%.

Relevante ting å diskutere:

- En mulig diskusjon kunne gå på om SØVN medierer effekten av alder. Det er uansett fortsatt en direkte negativ effekt av alder, som ikke kan forklares av SØVN.
2. Dersom du var ansvarlig for en studie der målet var å avdekke en mulig sammenheng mellom verbal hukommelse og søvn, gjør rede for hvordan du ville gå frem for å sørge for at du ville være i stand til å påvise en slik sammenheng (hint. hvilke faktorer påvirker prediktorens signifikans og gyldighet).

Kjernepoeng som bør komme frem:

- Faktorer som påvirker prediktorens signifikans er de som påvirker dens standardfeil. En kort definisjon av standardfeil kan med fordel gis.
 - Kandidaten er forventet å kjenne til at følgende uttrykk for estimert standardfeil i en bivariat regresjon $SE(\hat{b}_1) = \frac{s}{\sqrt{\sum(X_i - \bar{X})^2}}$ og at det gjennom den er tre forhold som påvirker standardfeilen (og følgelig signifikans) til en uavhengig variabel. Dette er:
 1. Antall observasjoner. Vi må sørge for at vi har nok observasjoner, da dette gir større statistisk styrke. Standardfeilen får ned ettersom antall observasjoner øker (summen i nevneren blir større).
 2. Når «s» i telleren er liten blir standardfeilen mindre. Dette kan skje blant annet ved:
 - a. Innflytelsen av tilfeldig målefeil minimeres (bruk et reliabelt mål på verbal hukommelse).
 - b. Andre variabler (selv om de er ukorrelerte med de andre prediktorene) står nok også for en del av den uforklarte variansen rundt regresjonslinjen («s»), og dersom de inkluderes i modellen vil s synke. Dette vil også gi lavere p-verdier.
 3. Vi er avhengig av betydelig varians i X (SØVN). Vi bør inkludere deltagere med stor variasjon i hvor mye de sover. Da vil nevneren også bli større, og standardfeilen mindre.
- I tillegg til «signifikans» spørres det etter «gyldighet». Her bør det nevnes at andre mulige konfunderende variabler bør inkluderes, for ellers risikerer vi å feilaktig tilskrive effekt til SØVN som egentlig kausalt skyldes andre forhold (slik som i oppgaven over).

Andre ting som kan være relevant å nevne:

- Modelltilpasning (hvordan finne det beste subsettet av prediktorer). F.eks. forward/backward elimination. Informasjonskriterier, kryssvalidering.

d) I modell 4 legges følgende uavhengige variabler til; FYSISK, SOSIAL og HODE.

- Hvilke av de uavhengige variablene i modell 4 er statistisk signifikant assosiert med VERBAL på et 0,05-nivå? (Begrunn svaret)
 - SØVN.C: **Signifikant**, $p < 0.05$. Dette flagges også av R gjennom de to stjernene «**», som antyder at $p < 0.01$. Nøyaktig p-verdi er 0.007157.
 - ABSTRAKT: **Signifikant**. Her bruker vi det faktum at siden utvalget er så stort, vil den nøyaktige samplingfordelingen til t-statistikken hunder H_0 være t-fordelt med $n-p-1$ frihetsgrader. Utifra utskriften bør kandidaten se at utvalget består av 400 individer, og fordelingen i dette tilfellet er derfor $t(400-7-1)=t(392)$. Denne fordelingen vet vi er tilnærmet identisk med en standard normalfordeling (z-fordeling), så vi kan bruke de kritiske verdiene til denne (+2). Siden t-verdien er $\ll -2$ (-37.375), er $p < 0.05$.
 - SØVN.CxABSTRAKT: **Signifikant**. Her må vi regne ut t-verdien og bruke resonnering over. $t = 0.108071 / 0.029880 = 3.616834$. p-verdien er følgelig < 0.05 , men vesentlig høyere enn for ABSTRAKT.
 - ALDER: **Signifikant**. Her må vi se på konfidensintervallet. En kort definisjon av hva er konfidensintervall er viser dypere forståelse. Et (1-alfa) konfidensintervall består av de verdiene for parameteret vi ikke kan forkaste på et alfa nivå. Verdier utenfor intervallet kan vi forkaste. Siden den øverste grensen for CI for ALDER er < 0 , betyr dette at 0 ligger utenfor intervallet, utgjør ikke en plausibel populasjonsverdi, og nullhypotesen kan forkastes.
 - FYSISK: **Signifikant**. Vi kan resonnerer tilsvarende som ALDER. Nedre grense for intervallet er > 0 , og følgelig ligger 0 utenfor intervallet.
 - SOSIAL: **Ikke signifikant**. Vi ser at t-verdien er i intervallet $[-2,2]$. Observert verdi på teststatistikken er derfor mindre enn den kritiske verdien.
 - HODEOMKRETS: **Ikke signifikant**. Intervallet rommer verdien 0.
- Forklar hvorfor forholdet mellom de standardiserte koeffisientene for variablene FYSISK og SOSIAL er så forskjellig fra forholdet mellom de ustandardiserte koeffisientene.

Her bør kandidaten vise at han/hun forstår hva en standardisert regresjonskoeffisient er, og hvorfor vi kunne være interessert i å sammenlikne de to koeffisientene. Tolkningen til en standardisert koeffisient er «antall standardavvik endring i forventet verdi på den avhengige variabelen som følger av en endring lik ett standardavvik på den uavhengige». Dette lar oss sette variablene på en felles skala, hvilket i større grad muliggjør sammenlikning av dem. Av betydning her er at den ustandardiserte koeffisienten til FYSISK er mindre enn SOSIAL. Dette følger derimot av at FYSISK er mål i enheter av minutter, mens SOSIAL har enhet timer. En enhet SOSIAL er derfor trolig en mye større endring enn en enhet FYSISK, og den ustandardiserte koeffisienten vil også bare som følge av dette bli mindre.

e) Vi skriver ofte uttrykket for en bivariat regresjonsmodell slik:

$$y_i = b_0 + b_1 X_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

- Forklar hvilke antagelser du leser ut av «epsilon» (ϵ_i) i uttrykket over.

Flere forhold kan påpekes. Uttrykket over spesifiserer modellen vi har for den avhengige variabelen i populasjonen. Observert skåre for et gitt individ «i» er en lineær funksjon av personens skåre på den uavhengige variabelen (X_i), pluss et avvik fra forventet verdi. Avviket (error, epsilon) er en feilterm som skyldes det samlede bidraget til målefeil og alle andre umålte forhold. Vi modellerer dette avviket som en tilfeldig (stokastisk) verdi trukket fra en

normalfordeling med snitt 0 og en gitt varians). Av dette følger det at avviket fra forventet skåre skal være normalfordelt, og siden alle avvikene er trukket fra samme normalfordeling (ingen subskript «i» i σ^2), så skal variansen være lik for alle nivåer av X. Følgende forventninger følger direkte av uttrykket over:

- Et histogram av residualene skal være tilnærmet normalfordelt.
- Variansen av residualene skal være tilnærmet konstant, og ikke avhengig av observerte skåre på de uavhengige variablene (homoskedastisitet).
- Ekstreme verdier (uteliggere) bør ikke inntreffe, da det er ekstremt usannsynlig å trekke slike verdier fra den tenkte normalfordelingen.
- Feiltermene (og følgelig residualene) skal være uavhengige. Hver epsilon er trukket fra samme fordeling, helt uavhengig av verdiene på de andre epsilonene.

2. Under utskriften fra regresjonsanalysene finner du et scatterplot av residualer fra en regresjonsanalyse. Hva ser du i plottet som du tenker kan utgjøre et problem for analysene, og hvorfor kan det bli et problem?

I figuren har vi ganske tydelig en enkelt uteligger, dvs. Y-verdien for en observasjon ligger langt fra de øvrige (gjørne spesifisere at den er $>1.5 \cdot$ den interkvartile avstanden). Det er ingen uteliggere i X-verdiene, og utover den enkle observasjonen ser vi ingen tydelige brudd på antagelsene.

Uteliggere i Y er problematiske fordi statistikker som benyttes i regresjon ikke er robuste, og kan sterkt påvirkes av slike verdier. Det er vanlig å benytte minste kvadraters metode for å finne estimater for b_0 og b_1 . Dette betyr at vi velger estimater av regresjonskoeffisientene som minimerer funksjonen $SSE = \sum (y_i - (\hat{b}_0 + \hat{b}_1 X_i))^2$. Uteliggere i Y kan sterkt påvirke estimatene fordi residualene kvadreres, og uteliggere vil føre til et enormt bidrag til SSE, noe som trekker estimatene i deres retning.

I hvilken grad uteliggere påvirker estimatene av intercept og stigningsgrad vil avhenge av antall observasjoner. Her er det så mange at det neppe vil bli radikale endringer i den estimerte regresjonslinjen. Derimot vil uteliggeren bidra sterkt til SSE, slik at vi kan ende opp med vesentlig lavere forklart varians enn vi ville hatt uten dem. Dette vil også føre til større standardfeil på estimatene, og lavere p-verdier.

Andre ting som kan være relevant å nevne:

- Uteliggere i X vil få høy *leverage*, og de vil være veldig bestemmende for estimatene. De mest problematiske datapunktene er derfor de som er utligger både i X og Y verdien.
- Hva burde vi gjøre med datapunktet? (potensielt droppe det)
- Standardiserte eller studentized residualer som en måte å flagge dem.
- Andre influensstatistikker (Leverage, DFBETA, Cook's d).

Oppgave 2: Forhold som påvirker innsats ved problemløsning (35%)

En forsker ønsker å forstå hvilke forhold som påvirker hvor hardt vi prøver å løse en logisk oppgave. En gruppe deltagere på totalt $n=90$ personer ble randomisert inn i en av tre betingelser, med like mange i hver. Deltagere i gruppe A løste oppgaven alene. Gruppe B hadde en hund i rommet mens de løste oppgaven, mens det sammen med deltagere i gruppe C satt et annet menneske i rommet mens de jobbet med oppgaven.

Den logiske oppgaven hadde ingen løsning, men i denne studien er avhengig variabel hvor lenge deltagerne jobbet med problemet før de ga opp.

a)

1. Fyll ut de sladdede verdiene i tabellen, og tolk resultatene fra analysen.

```
> AOV.F1=aov(Minutter~factor(Condition), data=DAT)
> summary(AOV.F1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Condition)	2	84.87	42.43	11.77	3e-05 ***
Residuals	87	313.70	3.61		

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

P verdien er $\ll 0.05$, slik at vi kan forkaste H_0 . Nullhypotesen er at dataene i alle tre gruppene er trukket fra en (normal) fordeling med samme forventede verdi (gjennomsitt).

Altså er vi rimelig sikker på at de observerte forskjellene mellom de tre gruppene i antall minutter deltageren jobber med oppgaven ikke skyldes tilfeldig samplingvariasjon.

Besvarelsen må gjerne gi mer bakgrunn: Både MS_w og MS_b (eventuelt MSM og MSE) er forventningsrettede estimater på populasjonsvariansen, og forventede verdi på F-statistikken er derfor 1. Store verdier på F-statistikken taler imot H_0 . 11.77 er således langt over kritisk verdi for å kunne forkaste H_0 på alfa lik 0.05.

Det er relevant å følge dette opp med en diskusjon om at ANOVA er en omnibus test, og hvordan gå frem for å gjøre parvise sammenlikninger. Relevante temaer er planned-comparisons, post-hoc tester, korreksjon for multiple tester, familywise-error, effektstørrelse.

Andre ting som kan være relevant å nevne:

- Samplingfordelinger, standardfeil. Faktorer, nivåer.

2. Vurder om følgende påstand er sann, og begrunn svaret ditt;

- I en enveis ANOVA tolker vi en større innen-gruppe-variasjon enn mellom-gruppe-variasjon som bevis mot nullhypotesen.

«Variasjon» er et upresist begrep, og burde med fordel tolkes i besvarelsen. Forskjeller mellom deltagere innad i gruppene (SS_w) er per definisjon uforklart, mens forskjeller mellom gruppene (SS_b) kan fremkomme enten på grunn av reelle forskjeller mellom gruppene eller samplingvariasjon. Vi gjør ikke direkte antagelser om variasjon, men det ligger til grunn for ANOVA at under H_0 er både MS_w og MS_b forventningsrettede estimater på populasjonsvariansen. Ratioen MS_b/MS_w har forventet verdi 1, og store verdier taler mot H_0 . Vi forkaster altså H_0 når det er større estimert verdi på bakgrunn av forskjeller *mellom* gruppene enn forskjeller *innad* i gruppene, altså det motsatte av påstanden i oppgaven.

Slik det er formulert er påstanden er gal.

3. Under ANOVA utskriften finner du utskrift fra en regresjonsanalyse der vi også ser på sammenhengen mellom Betingelse og Minutter, og der "Alene" er referanse kategorien.

Sammenlikn resultatene med utskriften fra variansanalysen. Hvilke slutninger kan du trekke om Dummy1 og Dummy2 ut fra utskriften?

Viktig her at kandidaten kan gjøre rede for at både lineær regresjon og ANOVA er medlemmer i samme klasse, generelle lineære modeller. Gjennom F-verdien og tilhørende p-verdi, testes derfor samme nullhypotesen som i enveis ANOVA (og F-verdien er da den samme). Konklusjonen blir derfor også her at vi kan forkaste hypotesen om at dataene er samlet fra grupper med samme gjennomsnitt, dvs. vi tror altså ikke de observerte forskjellene i gjennomsnitt skyldes tilfeldighet.

Variabelen «Betingelse» er kategorisk, og for å inkludere en slik variabel som uavhengig variabel i modellen må den rekodes. Slik det er rekodet her, er betingelse 1 (alene) referansekategori (0 på begge dummy variablene).

$$\widehat{Minutter}_i = \hat{b}_0 + \hat{b}_1 Dummy1_i + \hat{b}_2 Dummy2_i$$

Siden personer i referansekategorien har 0 for både D1 og D2, blir \hat{b}_0 lik estimert verdi på antall minutter blant deltagerne som løser oppgaven alene. \hat{b}_1 blir følgelig estimert forskjell i antall minutter mellom betingelse 1 og betingelse 2. De som løser oppgaven sammen med en hund har forventet utholdenhetstid lik 0.164 min kortere enn de som sitter alene. T-verdien lar oss signifikanteste denne forskjellen, og vi kan ikke forkaste hypotesen om at de er like.

\hat{b}_2 forstås som differansen i snitt mellom gruppe 1 og gruppe 3. De som løser oppgaven i et rom sammen med en annen person holder ut 1.9730 minutter lengre før de gir opp. Denne forskjellen er statistisk signifikant.

Andre ting som kan være relevant å nevne:

- Betydning av koding av variabelen, kontraster, hvordan teste forskjellen mellom Gruppe 2 og Gruppe 3. Multipl testing

b) Halvparten av deltagerne satt ovenfor et stort speil mens de jobbet med oppgaven, en intervensjon som brukes for å øke deltagerens selvbevissthet.

- Vedlagt finner du utskrift fra en toveis variansanalyse med faktorene BETINGELSE og SPEIL. Tolk resultatene, og sammenlikn med resultatene fra enveis variansanalysen i oppgave 2a.

Utskriften viser en signifikant hovedeffekt av betingelse, hovedeffekt av speil og en interaksjon.

Det er altså troverdige forskjeller i tid brukt på tvers av betingelse (alene, hund, person). Vi ser ut ifra linjediagrammet at dette trolig skyldes at folk i «person» gruppen bruker lengre tid.

Det er reelle forskjeller mellom de som sitter foran et speil og de som ikke gjør det. Folk som løser oppgaven i et rom med et speil holder ut lengre enn de som ikke gjør det (turkis linje ligger over oransje i alle betingelser).

Det viktigste å bruke tid på å diskutere er interaksjonseffekten. Denne bør defineres, og kandidaten bør forklare hvordan den observeres i linjediagrammet, og hva den substansielle tolkningen blir i eksperimentet.

I forhold til resultatene fra 2a ser vi at F-verdien av betingelse er større, og p-verdien derfor mindre. Årsaken bør diskuteres. Dette skyldes i korthet det faktum at SSw blir mindre, siden mye av denne kan nå forstås som grunnet forskjeller i eksponering for speil.

Fordelen med 2-veis resultantene er derfor at vi nå har en rikere forståelse av betydningen av betingelse (den avhenger av den andre faktoren, speil), og vi har økt statistisk styrke (lavere p-verdi).

Eta² og betydningen av effektstørrelser bør diskuteres. I overkant av 20% av variabiliteten i Minutter kan tilskrives både betingelse og speil. Interaksjonen står for rundt 5%. Over halvparten i forskjeller i bruk av tid kan ikke forklares ut ifra faktorene i eksperimentet.

Andre ting som kan være relevant å nevne:

Hva er en hovedeffekt, 2-veis ANOVA som regresjonsmodell, partiell eta².

Oppgave 3: Tastatur og syn (10%)

I en studie av menneskelige faktorer ønsker man å vurdere om folk med nedsatt syn foretrekker andre tastaturer enn folk med normalt syn.

Vedlagt finner du resultater fra analyser av variablene, TASTATUR og SYN. Variabelen TASTATUR har tre nivåer (1: Normalt, 2: Høy kontrast farger, 3: Bakbelyste knapper). Variabelen SYN har to nivåer (1: Svaksynt, 2: Normalt syn).

1. Hva ville du konkludert med vedrørende forholdet mellom og syn og preferanse for tastaturene?
 - H₀ burde spesifiseres (TASTATUR og SYN er uavhengige variabler).
 - Antall frihetsgrader, og bakgrunn for den kan med fordel nevnes.

Konklusjonen på analysen er at vi kan forkaste hypotesen om at TASTATUR og SYN er uavhengige variabler.

- For å forstå forholdet mellom variablene, og hvorfor nullhypotesen kan forkastes, må vi se i tabellen av residualer.
 - Her kan det gjerne forklares hva et residual er, og hva et standardisert residual er.
 - Residualene antyder at det er vesentlig høyere andel av svaksynte foretrekker et høykontrast tastatur enn det vi ville forventet ved tilfeldighet dersom variablene var uavhengige. Tilsvarende er det vesentlig flere normalsynte foretrekker et ordinært tastatur. Det er ingen forskjell mellom de to gruppene i deres preferanser for bakbelyste taster.

Andre ting som kan være relevant å nevne:

- Utrekning av kji-kvadrat verdien, og denne som en kvantifisering av relativ avstand mellom observert og forventet frekvenstabell). Frihetsgrader