Thor Arnfinn Kleven
Institutt for pedagogikk
19.09.2013

Effektstørrelse

Tradisjonelt har *signifikanstesting* vært fremhevet som den viktigste statistiske analyseformen i pedagogisk og psykologisk forskning. I de senere år har det blitt mye større oppmerksomhet rundt det som kalles *effektstørrelse*. En viktig grunn til dette er at statistisk signifikans betyr ikke nødvendigvis praktisk signifikans. Dersom man studerer et veldig stort utvalg, kan selv den minste sammenheng bli statistisk signifikant. Som eksempel: hvis N=1000, vil en Pearson r=0,09 være signifikant på 1 % nivå (se tabell 1). Vi vet imidlertid fra tidligere i kurset (se Howitt & Cramer 2011, s. 78) at r^2 uttrykker hvor stor andel av variansen som er fellesvarians for de to variablene. Når r= 0,09, er $r^2 < ,01$. altså mindre enn 1 %. I mange tilfeller er dette så lite at det har ingen praktisk betydning.

Tabell 1. Kritiske verdier for Pearson's produkt-moment-korrelasjon med 5% og 1% signifikansnivå.

N	5%	1%	N	5%	1%
10	.632	.765	100	.195	.256
20	.444	.561	200	.138	.181
30	.361	.463	300	.113	.148
40	.312	.403	400	.098	.128
50	.279	.361	500	.088	.115
60	.254	.330	600	.080	.105
70	.235	.306	700	.074	.097
80	.220	.286	800	.070	.091
90	.207	.270	900	.065	.086
			1000	.062	.081

Vi ser imidlertid også av tabell 1 at hvis N=20, vil ikke en Pearson r=0,40 være signifikant på 5 % nivå, selv om r=0,40 uttrykker en mye sterkere sammenheng enn r=0,09. Dette forteller oss at korrelasjonskoeffisienten og signifikanstesten svarer på ulike spørsmål.

Korrelasjonsspørsmålet kan formuleres slik:

> Hvor stor grad av sammenheng ser det ut til å være mellom variablene X og Y?

Det tilsvarende signifikansspørsmålet kan formuleres slik:

> Hvor sikkert er det at det i det hele tatt er noen (ikke-tilfeldig) sammenheng mellom X og Y?

Det er selvfølgelig sammenheng mellom disse to spørsmålene, men svaret på signifikansspørsmålet vil avhenge både av korrelasjonens størrelse og av utvalgets størrelse. Derfor ønsker vi svar på begge spørsmålene.

Signifikanstesting forteller altså bare med hvilken sikkerhet vi kan si at det er sammenheng til stede i populasjonen og ikke bare i utvalget. Hvis man har et tilfeldig utvalg fra en konkret populasjon, kan man for eksempel signifikansteste om forskjellen mellom resultatene for kvinner og menn i utvalget gir grunn til å tro at det er forskjell mellom kvinner og menn også i populasjonen.

Selv om man ikke har et tilfeldig utvalg fra en konkret populasjon, er det likevel interessant å vite om den forskjellen som er funnet, er så stor at det er lite sannsynlig at den ville oppstå ved ren tilfeldighet. Den populasjonen man signifikanstester i forhold til ved en slik signifikanstest, er en hypotetisk populasjon som man ikke kjenner. I et slikt tilfelle vil altså en signifikanstest ikke ha noe å gjøre med generalisering til en nærmere bestemt populasjon som vi ønsker å uttale oss om, men man får svar på hvor sannsynlig det er at en forskjell av en slik størrelse ville kunne oppstå ved ren tilfeldighet.

Selvfølgelig er det slik at jo større en sammenheng er, jo mindre sannsynlig er det at den kunne ha oppstått ved ren tilfeldighet. Men når spørsmålet om tilfeldighet skal vurderes, må man også ta i betraktning størrelsen av utvalget, ikke bare størrelsen av sammenhengen. En meget liten sammenheng kan altså være statistisk signifikant i et tilstrekkelig stort utvalg, mens på den annen side en tilsynelatende ganske sterk sammenheng kan oppstå ved ren tilfeldighet i et lite utvalg.

Hvis en sammenheng er ikke-signifikant, er vi tilbakeholdne med å tillegge den betydning. Det bør vi være, ettersom det er relativt stor sjanse for at en slik sammenheng er tilfeldig og derfor ikke er verd å gi noen tolkning. På den annen side følger det av signifikanstestingslogikken at nullhypotesen ikke kan bevises. Vi starter med å anta at nullhypotesen er riktig, og om vi finner at data stemmer godt med den antagelsen, beviser ikke det at antagelsen er riktig. Hvis data derimot ikke stemmer med

antagelsen, har vi ført et sannsynlighetsbevis for at antagelsen var feil, og konkluderer med signifikans.

Et ikke-signifikant resultat betyr altså ikke at nullhypotesen er bevist. Det betyr bare at vi ikke har funnet grunnlag for å forkaste nullhypotesen. Dersom samme tendens til sammenheng holder seg i et dobbelt så stort utvalg, vil kanskje sammenhengen være signifikant. Derfor kan størrelsen av sammenhengen i et utvalg ha interesse selv om sammenhengen er ikke-signifikant. Samtidig må man ha klart for seg, og gjøre det klart også for leseren, at så lenge dette bare er undersøkt i et så lite utvalg, kan man ikke ha tillit til at resultatet uttrykker noe ikke-tilfeldig. Hvis vi for eksempel tenker på en behandlingsform for en sjelden lidelse, eller et lesetreningsopplegg for elever med en spesiell form for lesevansker, er det vanskelig å gjennomføre undersøkelser med store utvalg. Kanskje har man en behandlingsform som er utprøvd over lang tid, på mange personer, som har en statistisk signifikant effekt, men effekten er moderat. La oss så tenke oss at man prøver ut en ny behandlingsform som foreløpig bare er utprøvd på få personer. Denne nye behandlingen har hatt god effekt på de få som er undersøkt, men siden utvalget er lite, er effekten ikke statistisk signifikant. I en slik situasjon kan vi karakterisere den nye behandlingen som mer lovende, men så lenge den ikke er utprøvd på flere personer, kan vi ikke påstå at den har en virkning. Vi er sikrere på at den gamle behandlingen virker, selv om effekten av denne er moderat.

Så langt i dette notatet er det for det meste brukt uttrykket sammenheng, mens uttrykkene effekt og effektstørrelse er brukt i overskriften og i det forrige avsnittet. Det er viktig å være oppmerksom på at i statistikklitteraturen og i forskningsrapporter brukes uttrykket effektstørrelse om statistiske sammenhenger, selv om de ikke nødvendigvis forteller noe om et årsakforhold mellom variablene. Det er egentlig litt uheldig, for ordet effekt uttrykker en årsakssammenheng. I et kontrollert eksperiment er det rimelig å tolke forskjellen mellom eksperimentgruppe og kontrollgruppe som et uttrykk for eksperimentvariabelens effekt. I andre typer undersøkelser bør man være forsiktig med å tale om effekt før spørsmål som berører indre validitet, er vurdert.

I dette notatet brukes altså betegnelsene grad av sammenheng og effektstørrelse i samme betydning. Grunnen til det er at uttrykket effektstørrelse i statistikklitteraturen simpelthen betyr grad av sammenheng. I resten av notatet vil det likevel bli kalt effektstørrelse, siden det er det vanlige uttrykket i litteraturen.

At effektstørrelse simpelthen betyr grad av sammenheng ser vi også ved at de fleste effektstørrelsemålene er basert på korrelasjoner. Alle korrelasjonskoeffisienter uttrykker grad av sammenheng, og kan følgelig benyttes som mål for ”effektstørrelse”. Men først skal vi likevel se på et effektstørrelsemål som er utviklet i en eksperimentell sammenheng, og som er basert på størrelsen av forskjellen mellom to gjennomsnittsverdier.

Effektstørrelse knyttet til forskjell mellom to grupper

Effektstørrelse knyttet til forskjell mellom to grupper kan estimeres enten på grunnlag av forskjellen mellom gruppegjennomsnittene eller på grunnlag av korrelasjon mellom gruppetilhørighet og resultat på den aktuelle variabelen.

Tabell 2. Dataeksempel for utregning av Cohen's d.

	N	Gj.snitt	St.avvik	t	p-verdi
Eksp.gruppe	50	20,42	3,16	2,44	0,02
Kontrollgruppe	50	18,66	4,00		

I eksemplet som er vist i tabell 2 tenker vi oss en eksperimentell undersøkelse der en eksperimentgruppe har fått en påvirkning som kontrollgruppen ikke har fått. Ved måling av avhengig variabel etter tiltaksperioden ble resultatene som vist i tabellen. Forskjellen mellom eksperimentgruppe og kontrollgruppe er signifikanstestet med t-test, og tabellen viser at forskjellen er signifikant med p-verdi 0,02.

Effektstørrelsemålet **Cohen's d** tar utgangspunkt i differansen mellom gjennomsnittene, og uttrykker denne med standardavviket som måleenhet, altså

$$Cohen's\ d = \frac{\bar{x}_{e-}\bar{x}_k}{s}.$$

I eksperimentelle undersøkelser har det vært argumentert for å dividere på standardavviket for kontrollgruppen, men det vanligste er at man dividerer på et veid gjennomsnitt av standardavvikene i gruppene, etter følgende formel:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}.$$

I talleksemplet blir s=3,6, og Cohen's d = (20,42-18,66)/3,6 = 0,49.

Til hjelp for tolkingen har Cohen laget følgende tommelfingerregel for Cohen's d:

0,2 – small 0,5 – moderate 0,8 large

I eksemplet over kan man altså å si at effektstørrelsen er moderat. Men slike tommelfingerregler er nokså vilkårlige, og bør brukes med stor forsiktighet. Det er vanligvis mer fornuftig, og også mer informativt, å tolke effektstørrelse i relasjon til standardavvik enn å fokusere på verbale beskrivelser som liten, moderat eller stor. Siden standardavviket er brukt som måleenhet ved utregning av Cohen's d, kan resultatet i eksemplet tolkes som at gjennomsnittsresultatet for eksperimentgruppen er cirka et halvt standardavvik høyere enn gjennomsnittsresultatet for kontrollgruppen.

Et alternativt effektstørrelsemål er korrelasjonen mellom eksperimentvariabelen og avhengig variabel. Dersom vi gir alle i eksperimentgruppen verdien 1 på eksperiment-variabelen og alle i kontrollgruppen får verdien 0, kan eksperimentvariabelen korreleres med avhengig variabel. Korrelasjonen mellom en dikotom variabel og en kontinuerlig variabel kalles **punkt-biserial korrelasjon**, men korrelasjonskoeffisienten kan regnes ut som en vanlig Pearson r. I talleksemplet ovenfor blir r=0,239, og r^2=0,057. Både r og r^2 kan tolkes som effektstørrelsemål. Den mest konkrete tolkningen får vi ved hjelp av r^2, siden r^2=0,057 forteller at 5,7 % av variansen i avhengig variabel kan forklares ved hjelp av eksperimentvariabelen.

I eksperimentelle undersøkelser har man for det meste benyttet Cohen's d som effektstørrelsemål, mens korrelasjoner og kvadrerte korrelasjoner har blitt brukt i ikke-eksperimentelle studier. Dette har imidlertid bare med tradisjoner å gjøre. Både Cohen's d og r og r^2 er like anvendbare i eksemplet ovenfor, og det samme gjelder om vi tenker at grupperesultatene i tabellen ovenfor gjelder jenter og gutter i stedet for eksperimentgruppe og kontrollgruppe.

Det er viktig å være oppmerksom på at tallstørrelsene for Cohen's d og korrelasjon ikke er direkte sammenlignbare. Hvis man vil ha en tilsvarende tommelfingerregel for hva som er liten, moderat og sterk effekt, målt med korrelasjon, sier man at r=.10 uttrykker en svak effekt, r=.30 er en moderat effekt, mens r=.50 uttrykker en sterk effekt. Det gir imidlertid mer informasjon å bruke r^2 til å tolke korrelasjonen som forklart varians, og da vil r=.10 innebære at 1 % av variansen i avhengig variabel kan forklares av uavhengig variabel, mens de tilsvarende tallene for r=.30 og r=.50 er henholdsvis 9 % og 25 %.

Effektstørrelse knyttet til forskjell mellom flere grupper

Tabell 3. Resultater fra variansanalyse av resultater på en leseprøve ved tre ulike skoler. (Utregning av en slik tabell er vist i Howitt & Cramer, kap. 19-20.)

Variasjonskilde	Kvadratsum	Frihetsgrader	Varians	F	p (sig.)
Mellom skoler	35,07	2	17,54	4,739	.009
Innom skoler	1032,42	279	3, 70		
Total	1067,49	281			

Anta at tabell 3 viser en variansanslysetabell for elevresultatene på en leseprøve som har blitt tatt ved tre ulike skoler. F-testen viser signifikans (p<.01), så det er altså grunn til å tro at forskjellen i resultat mellom skolene ikke bare har med tilfeldigheter å gjøre.

eta og **eta²** er korrelasjonsbaserte effektstørrelsemål som uttrykker hvor stor denne forskjellen mellom skolene ser ut til å være. Vi finner eta^2 fra variansanalysetabellen over ved simpelthen å dividere kvadratsum mellom skoler på total kvadratsum:

$$eta^2 = \frac{SS_b}{SS_{tot}}$$

I talleksemplet blir eta^2=0,033 og eta=0,181.

Vi kan forestille oss eta som Pearson r mellom to tallrekker, der den ene rekken inneholder elevenes individuelle resultater og den andre rekken inneholder gjennomsnittsresultatet for den skolen eleven hører hjemme i. Dette medfører at vi kan tolke eta^2 som forklart varians i avhengig variabel. Når eta^2 i dette tilfellet er lik 0,033, betyr altså det at 3,3 % av variansen i resultat på leseprøven kan forklares ut fra hvilken skole eleven går på.

Det er verdt å merke seg at akkurat som t-test for differanse mellom to grupper kan betraktes som et spesialtilfelle av F-test, kan også punkt-biserial korrelasjon betraktes som et spesialtilfelle av eta-korrelasjon.

Korrelasjonskoeffisienter som effektstørrelsemål

Alle korrelasjonskoeffisienter uttrykker grad av sammenheng, og kan følgelig brukes som "effektstørrelsemål".

Når det gjelder sammenheng mellom *to kontinuerlige variabler*, er naturligvis Pearson r den mest aktuelle korrelasjonskoeffisienten.

I det foregående har vi sett at eta-korrelasjon kan brukes for sammenheng med en *nominalvariabel* (for eksempel skoler) og en *kontinuerlig variabel*. Dette gjelder enten nominalvariabelen har to eller flere nivåer, og hvis to nivåer er eta identisk med det som tradisjonelt kalles punktbiserial korrelasjon.

Det gjelder videre for både Pearson r og eta-korrelasjon at kvadrert korrelasjon kan tolkes som proporsjon forklart varians. eta^2 forteller hvor stor andel av variansen i avhengig variabel som kan forklares utfra gruppetilhørighet, mens r^2 uttrykker hvor stor andel av variansen i de kontinuerlige variablene som er felles for de to variablene.

Sammenheng mellom to variabler som begge er på *nominalnivå*, signifikanstestes med χ^2-test. Hvis variablene er dikotome (har bare to nivåer), er den tilhørende korrelasjonen en *phi-korrelasjon*, som kan finnes ved følgende formel (se Howitt & Cramer 2011, s. 422):

$$r_{phi} = \sqrt{\frac{\chi^2}{N}}.$$

Phi-korrelasjon er et effektstørrelsemål, men siden vi har med nominalvariabler å gjøre, gir det ingen mening å tolke kvadratet av denne korrelasjonen som forklart varians.

Hvis ikke nominalvariablene er dikotome, eller hvis variablene er på ordinalnivå, finnes det andre aktuelle korrelasjonsmål som kan uttrykke grad av sammenheng. Disse omhandles ikke i dette notatet, ettersom de heller ikke omtales i pensumlitteraturen.

Betydningen av å kombinere signifikanstest og effektstørrelsemål

I begynnelsen av dette notatet er det vist at en signifikanstest og en korrelasjonskoeffisient svarer på ulike spørsmål. Leseren får derfor mer informasjon når begge oppgis.

Hvis vi for eksempel får vite at korrelasjonen mellom variabel X og Y er r=0,30 (r^2=0,09), og at korrelasjonen er signifikant på 5 % nivå, så vet vi at variablene ser ut til å ha 9 % felles varians, og at sjansen for at dette skal være tifeldig er mindre enn 5 %.

Tabell 1 i dette notatet viser kritiske verdier for Pearson r for signifikansnivåene 5% og 1%. I lærebøkene har vi slike tabeller over kritiske verdier for ulike signifikanstester. Ved hjelp av slike tabeller kan vi få greie på om p-verdien (sannsynligheten) er større eller mindre enn det signifikansnivået som er valgt. Med et dataprogram for statistisk analyse kan vi få sannsynligheten for at et resultat kan være oppstått ved tilfeldighet oppgitt mer nøyaktig. I stedet for at p<.05, kan vi for eksempel få vite at p=.028. Hvis vi har et effektstørrelsemål og p-verdien som viser hvor sannsynlig det er å finne et effektstørrelsemål av denne størrelse eller større ved ren tilfeldighet, har vi fått det vi vanligvis trenger å vite om det tallmessige resultatet.

Effektstørrelse og meta-analyse

Meta-analyse (Howitt & Cramer, kap. 35) er statistisk analyse som sammenfatter resultater fra mange undersøkelser av samme fenomen. Dette er relevant når man skal summere opp hva forskningen så langt har vist om en problemstilling. Tradisjonelt har forskerne brukt sitt beste skjønn og summert opp sine litteraturstudier uten bruk av statistiske analyser. Utviklingen av meta-analyse har gitt nye muligheter for mer presise oppsummeringer, der man for eksempel kan estimere et effektstørrelsemål som er basert på resultater fra mange undersøkelser av det samme fenomen.

Avslutningsvis skal dette notatet vise hvorfor meta-analyse må baseres på opplysninger om effektstørrelse og ikke på signifikanstester fra de enkelte studiene.

Hvis vi har funnet en del undersøkelser som har studert samme fenomen, kunne det være en nærliggende tanke å se hvor mange av disse undersøkelsene som har funnet en signifikant sammenheng mellom de variablene som inngår i problemstillingen, og hvor mange som ikke fant signifikans. Et enkelt eksempel kan illustrere hvorfor st slikt resonnement ville kunne bære galt av sted.

Tenk deg at to forskere uavhengig av hverandre studerer korrelasjonen mellom X og Y. Begge har et utvalg på N=50, og begge finner at r= 0,25. I følge tabell 1 i dette notatet er r=0,25 ikke signifikant på 5 % nivå. Det foreligger altså to undersøkelser som begge har konkludert med at det ikke er signifikant sammenheng mellom X og Y, og man kan stå i fare for å tenke at her har vi to undersøkelser som bekrefter hverandre på at det ikke er noen sammenheng mellom variablene.

Men det relevante sannsynlighetsspørsmålet når undersøkelsene ses under ett, er: Hvor sannsynlig er det at to undersøkelser uavhengig av hverandre skulle komme så nær signifikansnivået som disse to har gjort? Sannsynligheten for det er mindre enn 5 %. Til sammen har de to forskerne studert 100 personer, og hvis disse hadde vært i samme utvalg, ville r=0,25 ha vært signifikant, i følge tabell 1. Men hver av de to forskerne hadde for lite utvalg til at sammenhengen kunne sies å være signifikant.

Vi ser altså at om vi kombinerer resultater fra to undersøkelser som hver for seg har vært ute av stand til å si at resultatet skyldes mer enn tilfeldigheter, så kan konklusjonen bli at det sannsynligvis likevel er en ikke-tilfeldig sammenheng mellom variablene.

I meta-analyse kombinerer man derfor *ikke* sannsynlighetsresonnement fra ulike undersøkelser, men effektstørrelser. Howitt & Cramer viser prinsipper for meta-analyse i kap. 35. Det ligger imidlertid utenfor målsettingen for PED4010 å kunne gjennomføre meta-analyse.