Kjikvadrattesten

Silje Synnøve Lyder Hermansen

s.s.l.hermansen@stv.uio.no

Effekt av sosial bakgrunn for lønnsforventning blant studenter

Univariat analyse

Gjennomsnittlig forventet lønn blant studenter ved HiØ er 614478. Hva kan forklare dette?

- H_1 Studenter som har foreldre med høy utdannelse forventer seg høyere lønnsnivå.
- H_2 Studenter som studerer realfag forventer seg høyere lønnsnivå.
- H_3 Studenter med foreldre med høy utdannelse velger realfag. Derfor forventer de seg høyere lønn.

Bivariat analyse – gjennomsnittsverdier i krysstabell

Gjennomsnittlig forventet lønn blant studenter ved HiØ er 614478. Hva kan forklare dette?

- H_1 Studenter som har foreldre med høy utdannelse forventer seg høyere lønnsnivå.
- H_2 Studenter som studerer realfag forventer seg høyere lønnsnivå.

	Nei	Ja
Studerer realfag	521438	776190
Foreldres utdannelse	473556	705071

Vi finner at studenter som studerer realfag forventer seg 254752 mer i lønn enn andre studenter. Tilsvarende ser vi at studenter fra bedrestilte hjem forventer seg 231516 kr mer i året.

Bivariat analyse – Pearsons R

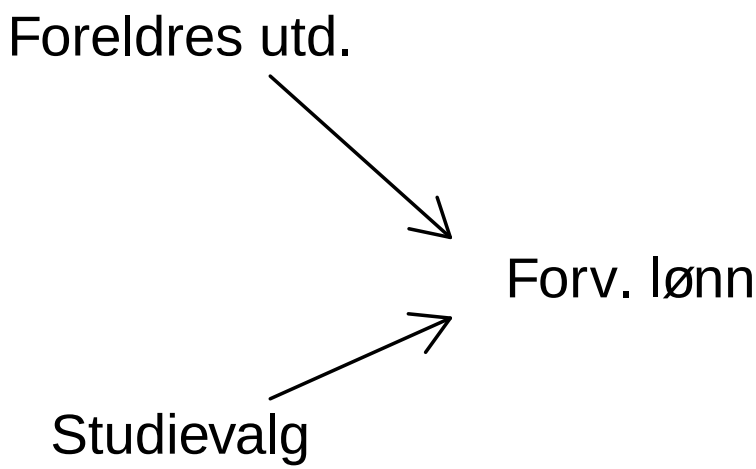
Med en tellevariabel (Lønn) og to binære variabler (Foreldres utdanning og Studievalg), kan jeg gjøre en korrelasjonsanalyse for metriske variabler. Jeg gjør en Pearsons R:

	Pearson	R2
Foreldres utdanning og lønn	0.31	0.10
Studerer realfag og lønn	0.33	0.11
Realfag og foreldresutd.	0.09	0.01

Foreldres utdanning forklarer omtrent 10 prosent av forventet lønn. Studievalg forklarer omtrent 11 prosent av forventet lønn.

Bivariat analyse – gjennomsnittsverdier i krysstabell

Foreløpig har vi testet to separate årsakssamenhenger:



Trivariatanalyse – årsaksmodell

Vi tester hypotese 3 om en spuriøs effekt mellom studievalg og forventet lønn.

- H_3 Studenter med foreldre med høy utdannelse velger realfag. Derfor forventer de seg høyere lønn.

Allerede fra de lave verdiene til korrelasjonen mellom studievalg og sosialbakgrunn kan vi slutte at det er lite mulighet for en spuriøs effekt mellom lønn og studier. Men la oss forsøke likevel.

Trivariatanalyse – årsaksmodell

Trivariat tabell

Vi ser at barn av høyt utdannede foreldre forventer seg høyere lønn enn andre, selv med likt kompetansenivå. Faktisk forventer de seg høyere lønn enn alle, uansett studieretning.

	Studerer ikke realfag	Studerer realfag
Foreldre uten høyere utd.	435806	557143
Foreldre med høyere utd.	584643	885714

Table : Gjennomsnittlig forventet lønn blant studenter ved HiØ. Trivariat tabell.

Trivariat tabell – tolkning

Vi kan sjekke den relative endringen i lønnsforventning for studentene. Her her den forventede lønnsandelen i forhold til gjennomsnittet blant respondentene:

	Bivariat	Studerer ikke realfag	Studerer realfag
Foreldres utd.: Høy	1.15	0.95	1.44
Studievalg: Realfag	1.26	0.91	1.44

Table : Relativ forventning til lønn

Trivariat tabell – tolkning

- Studenter med foreldre med høy utdanning forventer seg 15 prosent høyere lønn enn studenter som har foreldre med lav utdanning.
- Effekten så godt som forsvinner når vi vurderer studieretningen.
- Studenter som studerer realvag forventer seg i gjennomsnitt 26 prosent høyere lønn enn snittet blant respondentene.
- Denne effekten *øker* til 44 prosent når vi tar hensyn til foreldres utdanningsnivå.
- Derimot forventer studenter som ikke studerer realfag, seg lavere lønn enn snittet. De forventer -9 prosent lavere lønn. Effekten er noe lavere blant studenter med høyt utdannede foreldre: De forventer -5 prosent lavere lønn.

Kjikvadrattesten
lønnsforventning blant studenter

For å hypoteseteste bør vi følge oppskriften:

- Velg test: For krysstabeller (kategoriske variabler) er kjikvadrattesten god. For gjennomsnitt (metriske variabler), bruker vi t-testen.
- Velg sannsynlighetsnivå: Hvor ofte vil jeg ta feil? Standard er 95 prosents konfidensintervall; det vil si at jeg åpner for å ta feil i 5 prosent av tilfellene.
- Beregne sannsynlighet: Jeg leser av tabellen for kjikvadratskårer eller t-skårer for å finne toleranseverdien min. Alle verdier mer ekstreme enn den, betyr at jeg forkaster nullhypotesen min og antar at resultatet ikke er tilfeldig.

Kjikvadrattest – velge test

Kjikvadrattesten er laget for krysstabeller. For kategoriske variabler er krysstabeller den eneste måten å gjøre bivariat statistikk.

	Høy utd.	Lav utd.	Total
Høy lønn	29	5	34
Middels lønn	11	7	18
Lav lønn	20	27	47
Total	60	39	99

I dette eksempelet er variablene mine kodet til å være kategoriske. I spørreundersøkelsen om forventet lønn blant studenter har jeg kategorisert svarene i tre kategorier: Høy lønn, middels og lav lønn. Er lønnsforventningen tilfeldig fordelt, eller avhenger de av hvorvidt foreldrene har høyere utdannelse (binær variabel)?

Kjikvadrattest – velge sannsynlighetsnivå

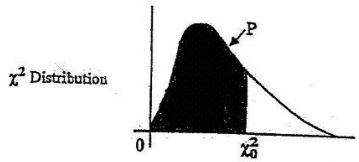
Jeg ønsker å ta feil i maks 5 prosent av tilfellene. Da må jeg regne ut antall frihetsgrader. I krysstabeller tar ikke testen hensyn til antall enheter i utvalget, men antall verdikombinasjoner (ruter i tabellen). For å beregne frihetsgradene ganger vi antall linjer i tabellen minus én med antall kolonner minus én:

$$df = (r - 1)(k - 1) \tag{1}$$

$$df = (3 - 1)(2 - 1) \tag{2}$$

$$df = 2 \tag{3}$$

Kjikvadrattest – velge sannsynlighetsnivå



The table below gives the value x_0^2 for which $P[x^2 < x_0^2] =$ P for a given number of degrees of freedom and a given value of P.

Degrees of Freedom	Values of P									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

Kjikvadrattest – velge sannsynlighetsnivå

Jeg leser av tabellen: i kolonnen med sannsynlighetsnivå på 0.95 (jeg vil ha rett i 95 prosent av tilfellene) finner jeg linjen for 2 frihetsgrader. Her finner jeg den kritiske kjikvadrat-skåren min. Min kritiske verdi er 5,991. Alle verdier mer ekstreme enn den, vil bety signifikante resultater.

Kjikvadrattest – beregne sannsynlighetsnivå

Formelen for kjikvadrat er som følger:

$$\chi^2 = \Sigma \frac{(Observert - forventet)^2}{Forventet} \tag{4}$$

Det den i praksis gjør er å beregne forkjsellen mellom hva vi skulle forventet oss når alt er tilfeldig og hva vi faktisk observerte. Siden beregner testen hvor (u)sannsynlig dette utfallet er.

Kjikvadrattest – observerte frekvenser

Hvis jeg skulle respektert de univariate fordelingene mine (hvor mange som forventer høy, middels og lav lønn og hvor mange med utdannede foreldre), hva slags verdikombinasjoner skal jeg da se?

	Høy utd.	Lav utd.	Total
Høy lønn	29.00	5.00	34.00
Middels lønn	11.00	7.00	18.00
Lav lønn	20.00	27.00	47.00
Total	60.00	39.00	99.00

Det beregner jeg ved å multiplisere kolonnenesummene med radsummene i tabellen og dele på antall enheter. Dette er nullhypotesen min!

$\frac{60 \times 34}{99} = 20,6$

Kjikvadrattest – forventede frekvenser (under nullhypotese)

Hvis jeg skulle respektert de univariate fordelingene mine (hvor mange som forventer høy, middels og lav lønn og hvor mange med utdannede foreldre), hva slags verdikombinasjoner skal jeg da se?

	Høy utd.	Lav utd.
Høy lønn	20.61	13.39
Middels lønn	10.91	7.09
Lav lønn	28.48	18.52

Det beregner jeg ved å multiplisere kolonnenesummene med radsummene i tabellen og dele på antall enheter. Dette er nullhypotesen min!

$$\frac{60 \times 34}{99} = 20,6 \tag{5}$$

Kjikvadrattest – differansen mellom observerte og forventede frekvenser

	Høy utd.	Lav utd.
Høy lønn	8.39	-8.39
Middels lønn	0.09	-0.09
Lav lønn	-8.48	8.48

$$\textit{Observert} - \textit{Forventet} \tag{6}$$

Kjikvadrattest – differansen mellom observerte og forventede frekvenser

	Høy utd.	Lav utd.
Høy lønn	70.46	70.46
Middels lønn	0.01	0.01
Lav lønn	71.99	71.99

Matematikere setter ting i annen når de ønsker å beregne distansen mellom punkter: Da gjør det ikke noe at noe av differansen er negativ og noe er positivt.

$$(Observert - Forventet)^2 \tag{7}$$

Kjikvadrattest – differansen normalisert

	Høy utd.	Lav utd.
Høy lønn	3.42	5.26
Middels lønn	0.00	0.00
Lav lønn	2.53	3.89

$$\frac{(\textit{Observert} - \textit{forventet})^2}{\textit{Forventet}} \tag{8}$$

(9)

For å gjøre resultatene sammenliknbare mellom rutene, deler vi resultatet på forventet frekvens.

Kjikvadrattest – differansen normalisert

$$\chi^2 = \Sigma \frac{(Observert - forventet)^2}{Forventet} \tag{10}$$

$$\chi^2 = 15.1 \tag{11}$$

Til sist summerer vi alle rutene i tabellen for å få kjikvadratskåren. Jo større skåren er, jo mindre sannsynlig er det at den observerte krysstabellen er oppstått ved en tilfeldighet.
Jeg husker at jeg hadde definert mitt kritiske nivå til 5,991. 15.1 er langt høyere enn det. Det er derfor ikke tilfeldig (med 95 prosent sikkerhet) at vi har observert en forskjell i lønnsforventning blant studenter med foreldre med hhv. høy og lav utdanning.