11. Grupperte data.

11.1. Modell-tilpasning.

Dersom du kaster en terning mange ganger, forventer du at terningen viser ”1”, ”2”, …, ”6” like mange ganger. Den *forventede* sannsynlighetsfordelingen er altså:

Utfall	1	2	3	4	5	6
Sannsynlighet	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Jeg kaster en terning 60 ganger (jeg tilstår: Jeg fikk Excel til å simulere terningkastene). Observert resultat sammen med forventet resultat er vist i tabellen nedenfor:

Utfall	1	2	3	4	5	6
Observert O_i	15	6	10	8	8	13
Forventet E_i	10	10	10	10	10	10

Som ventet er ikke de observerte resultatene helt lik de forventede. Men er avvikene så store at vi mistenker terningen (eller Excel) for ikke å gi helt tilfeldige utfall? Mer presist setter vi opp disse hypotesene:

H_0: De observerte dataene er et tilfeldig utvalg fra en populasjon der forventet antall $E_i = 10$ for alle gruppene.

H_1: Dataene er ikke forenlig med at forventet antall $E_i = 10$ for alle gruppene.

Vi tester med signifikansnivå $\alpha = 0.05$.

For å utføre denne testen, benytter vi setningen nedenfor:

Vi har et tilfeldig utvalg av observasjoner der hver observasjon kan grupperes i en og kun en av *n* grupper. For gruppe nr *i forventer* vi E_i observasjoner, mens vi *observerer* O_i. Dersom dataene er tilfeldig fordelt, vil

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

være tilnærmet kji-kvadrat-fordelt med $n-1$ frihetsgrader. Tilnærmingen er brukbar dersom $E_i \geq 5$ for alle gruppene.

For vårt lille eksperiment får vi:

$$\chi^2 = \frac{(15-10)^2}{10} + \frac{(6-10)^2}{10} + \cdots + \frac{(13-10)^2}{10} = \frac{25+16+0+4+4+9}{10} = \underline{5.8}.$$

Av en kji-kvadrat-tabell ser vi at med signifikansnivå $\alpha = 0.05$ og $6-1=5$ frihetsgrader finner vi en kritisk verdi $\chi^2 = 11.071$. Vår verdi for χ^2 ligger ikke i det kritiske området. Vi kan derfor ikke forkaste H_0. Dette medfører at vi ikke kan påstå at terningen (d.v.s. Excel)

ikke oppfører seg i henhold til de forventede verdiene. Men vi har heller ikke bevist at terningen virkelig oppfører seg som den skal.

Tilsvarende tester kan vi utføre i en lang rekke situasjoner der vi ut fra en eller annen modell *forventer* en bestemt fordeling, og vil sammenlikne denne med *observerte* data. Antall frihetsgrader er lik antall grupper minus antall bindinger som legges på de forventede verdiene. I eksemplet ovenfor har vi 6 grupper, og *en* binding som består i at summen av antall forventede verdier må være lik summen av antall observerte verdier. Derfor blir det $6-1=5$ frihetsgrader.

Eksempel 11.1: Vi skal nå se på antall ganger brannvesenet må rykke ut i løpet av en uke. Over et år (52 uker) observerer vi:

Antall utrykninger pr uke	0	1	2	3	4	5
Antall uker	6	21	14	5	4	2

Fordelingen minner mistenkelig om en Poisson-fordeling. Undersøk om dataene virkelig er Poisson-fordelte.

Løsning: For å kunne teste om dataene virkelig er Poisson-fordelt

$$P(x)=\frac{\mu^x\cdot e^{-\mu}}{x!},$$

må vi finne gjennomsnittsverdien μ:

$$\mu=0\cdot\frac{6}{52}+1\cdot\frac{21}{52}+2\cdot\frac{14}{52}+3\cdot\frac{5}{52}+4\cdot\frac{4}{52}+5\cdot\frac{2}{52}=\underline{1.73}.$$

Vi får da følgende tabell over forventet antall utrykninger pr uke:

X	0	1	2	3	4	≥ 5
$52\cdot P(x)$	9.22	15.95	13.80	7.96	3.44	1.63

For at kravet om at $E_i\geq 5$ for alle gruppene skal være oppfylt, må vi slå sammen de to siste gruppene. Vi får da denne tabellen:

X	0	1	2	3	≥ 4
O_i	6	21	14	5	6
E_i	9.22	15.95	13.80	7.96	5.07

Vi setter opp disse hypotesene:

H_0: Observert antall utrykninger pr uke stemmer med en Poisson-fordeling.

H_1: Observert antall utrykninger pr uke stemmer ikke med en Poisson-fordeling.

Vi tester med signifikansnivå $\alpha=0.05$.

Så setter vi i gang regningene:

$$\chi^2=\frac{(6-9.22)^2}{9.22}+\frac{(21-15.95)^2}{15.95}+\frac{(14-13.80)^2}{13.80}+\frac{(5-7.96)^2}{7.96}+\frac{(6-5.07)^2}{5.07}=\underline{3.998}.$$

Men hvor mange frihetsgrader har vi? Det er 5 grupper. Men vi har *to* bindinger: Den ene er at summen av alle E_i skal være lik 52. Den andre er at middelverdien skal være lik 1.73. Da får vi $5-2=3$ frihetsgrader. Kritisk verdi blir da $\chi^2 = 7.815$. Vår verdi ligger ikke i det kritiske området. Vi kan derfor ikke forkaste H_0. Men vi har heller ikke bevist at antall utrykninger pr uke virkelig er Poisson-fordelt.

11.2. Kontingenstabeller, uavhengighet.

Tabellen nedenfor viser sammenheng mellom ekteskapelig status og alkoholvaner for et tilfeldig utvalg amerikanere over 18 år:

		Alkoholvaner (drinker pr måned)			
		Avholdende	1 - 60	Over 60	Radsum
Ekteskapelig status	Enslig	67	213	74	354
	Gift	411	633	129	1173
	Enke/-mann	85	51	7	143
	Skilt	27	60	15	102
	Kolonnesum	590	957	225	1772

Denne tabellen er et eksempel på en ***kontingenstabell***, der vi ser to egenskaper i sammenheng. I tabellen over er det 4 rader med hver sin form for ekteskapelig status, og 3 kolonner med hver sin alkoholvane. Til sammen blir det 12 ulike kombinasjoner.

Vi skal nå undersøke om det er noen sammenheng mellom ekteskapelig status og alkoholvaner. Da setter vi opp disse hypotesene:

H_0: Det er ingen sammenheng mellom ekteskapelig status og alkoholvaner.

H_1: Det er en sammenheng mellom ekteskapelig status og alkoholvaner.

Vi tester med signifikansnivå $\alpha = 0.05$.

For å gjennomføre testen, må vi finne *forventet* antall personer i hver av de 12 rutene dersom H_0 er sann. La oss finne forventet antall i ruta for "Gift" og "Avholdende" som eksempel.
Av tallene i tabellen ser vi at:

- Sannsynligheten for at en tilfeldig person er "Gift" er $P(\text{Gift}) = \frac{1173}{1772}$.
- Sannsynligheten for at en tilfeldig person er "Avholdende" er $P(\text{Avholdende}) = \frac{590}{1772}$.

Dersom egenskapene "Gift" og "Avholdende" er uavhengige, har vi at

$$P(\text{Gift} \wedge \text{Avholdende}) = P(\text{Gift}) \cdot P(\text{Avholdende}) = \frac{1173}{1772} \cdot \frac{590}{1772}.$$

Forventet antall personer som er gift og avholdende blir da

$$E(\text{Gift} \wedge \text{Avholdende}) = 1772 \cdot \frac{1173}{1772} \cdot \frac{590}{1772} = \frac{1173 \cdot 590}{1772} = \underline{390.6}.$$

Tilsvarende resonnement kan vi gjennomføre for alle rutene. Vi får da denne generelle formelen:

Forventet antall i en rute i en kontingenstabell er

$$\frac{\text{Radsum} \cdot \text{Kolonnesum}}{\text{Totalsum}}.$$

Dersom vi gjennomfører disse beregningene i vårt eksempel, får vi denne tabellen over *forventede* antall (observerte antall i parentes):

		Alkoholvaner (drinker pr måned)			
		Avholdende	1 - 60	Over 60	Radsum
Ekteskapelig status	Enslig	117.9 (67)	191.2 (213)	44.9 (74)	354
	Gift	390.6 (411)	633.5 (633)	148.9 (129)	1173
	Enke/-mann	47.6 (85)	77.2 (51)	18.2 (7)	143
	Skilt	34.0 (27)	55.1 (60)	13.0 (15)	102
	Kolonnesum	590	957	225	1772

Så setter vi kalkulatoren i arbeid:

$$\chi^2 = \frac{(67-117.9)^2}{117.9} + \frac{(213-191.2)^2}{191.2} + \cdots + \frac{(15-13.0)^2}{13.0} = \underline{94.27}.$$

Men hvor mange frihetsgrader skal vi bruke? Her er det jo flere bindinger siden både radsummene, kolonnesummene og totalsummen skal stemme for de forventede antallene. Vi kan nå vise at:

I en kontingenstabell med ***r*** rader og ***c*** kolonner, er antall frihetsgrader

$$(r-1)(c-1).$$

I vårt eksempel får vi $(3-1)(4-1)=6$ frihetsgrader. Vi går inn i en kji-kvadrat-tabell, og finner en kritisk verdi på $\chi^2 = 12.592$. Vår observerte verdi ligger langt inne i det kritiske området, slik at vi forkaster H_0. Vi har altså påvist en sammenheng mellom ekteskapelig status og alkoholvaner.

I de tilfellene der vi forkaster nullhypotesen om uavhengighet, må vi være klar over:

- Vi har kun påvist at det med stor sannsynlighet eksisterer en sammenheng mellom to faktorer.
- Vi har *ikke* påvist hvilken sammenheng som eksisterer.
- Vi har kun påvist en *statistisk* sammenheng, ikke en *årsakssammenheng*.

Vær spesielt oppmerksom på det siste punktet. Selv om det kan påvises en statistisk sammenheng, kan denne sammenhengen skyldes faktorer som ikke inngår direkte i undersøkelsen, men som påvirker de egenskapene som inngår. I vårt eksempel kan bakenforliggende faktorer som alder, helse og sosial status påvirke både ekteskapelig status og alkoholvaner.