ECON 2130

HG
19. mars 2012

Supplement til forelesning 19. mars

Litt mer om den hypergeometriske fordelingen og dens tilnærming av binomisk fordeling.

Grunnen til dette supplementet er dels at jeg ikke rakk å gå igjennom alt jeg hadde planlagt på forelesningen, dels at forholdet mellom hypergeometrisk og binomisk fordeling er viktig og dels at Løvås ikke er helt korrekt i begynnelsen av avsnitt 5.3.

Hypergeometrisk fordeling:

1. Utgangspunktet er en populasjon bestående av N individer (ikke nødvendigvis personer). M av disse har et kjennetegn, A, mens resten, $N-M$, er $\overline{A}$.
2. Andelen av A i populasjonen er $p = M/N$.
3. Et utvalg på n individer trekkes *rent tilfeldig* fra populasjonen og uten tilbakelegging, der "*rent tilfeldig*" betyr at alle mulige utvalg på n har samme sjanse for å bli trukket ut. Et slikt utvalg kalles også ofte *et representativt utvalg*.
4. Det kan vises at kravet om *rent tilfeldig utvalg* er oppfylt dersom utvalget trekkes ved at ett og ett individ trekkes av gangen fra populasjonen slik at ved hver enkelt-trekning har alle gjenværende individer samme sannsynlighet for å bli trukket ut.
5. La X = antall A-er i utvalget.
6. Andelen av A i utvalget er, X/n, kan brukes til å estimere (anslå) $p = M/N$ dersom p er ukjent (mer om dette i kap. 6).

Hvis forutsetningen om rent tilfeldig utvalg er oppfylt, er X hypergeometrisk fordelt (kort skrevet: $X \sim \text{hypergeom}(N, M, n)$) med elementær sannsynlighetsfunksjon

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad \text{for } x = 0, 1, 2, \ldots, n$$

[**Merk** at, siden en binomialkoeffisient $\binom{r}{s} = 0$ når $s > r$, så kan noen av disse sannsynligheten være 0 (se nedenfor).]

Vi har dessuten (kan vises): $E(X) = np = n\frac{M}{N}$, $\text{var}(X) = np(1-p)\frac{N-n}{N-1}$

Hvis vi trekker utvalget ett og ett individ av gangen, kan hver trekning oppfattes som et forsøk, der vi i hvert forsøk registrerer om vi har en *A* (som vi kunne kalle *S*) eller $\overline{A}$ (som vi kunne kalle *F*). *X* blir dermed antall *S*-er i *n* forsøk (enkelt-trekninger). Under forutsetningen om rent tilfeldig utvalg (jfr. 3. og 4. ovenfor), gjelder

1) $P(S) = P(A) = M/N = p$ er konstant i alle forsøk (det er her Løvås tar feil)

2) I motsetning til binomiske forsøk er det her avhengighet mellom forsøkene (enkelt-trekningene) når det gjelder utfallet (S eller F). Denne avhengigheten blir (intuitivt sett) mindre og mindre ettersom populasjons-størrelsen, *N*, øker. Avhengigheten er neglisjerbar dersom *N* er stor nok i forhold til utvalgs-størrelsen, *n* (tommelfinger-regelen i Løvås, $N > 10n$, kan brukes).

Illustrasjon av 1) og 2) ved et regneeksempel

Vi trekker *n* = 13 kort fra en kortstokk på *N* = 52 kort. Kortene trekkes ett og ett av gangen. Det er *M* = 4 ess (*A*) i kortstokken, og *X* er antall ess i utvalget. Hvis kortstokken er godt stokket, kan vi anta at utvalget er rent tilfeldig, og hver enkelt-trekning er slik at alle gjenværende kort i kortstokken har samme sjanse for å trukket.

Vi skal først begrunne hvorfor 1) og 2) da er oppfylt: La E_j være begivenheten at vi trekker et ess i *j*-te trekning ($j = 1, 2, \ldots, 13$). Det er da klart at

(1) $P(E_1) = \frac{4}{52} = 0.0769, \qquad P(E_2 \mid E_1) = \frac{3}{51} = 0.0588, \qquad P(E_2 \mid \overline{E}_1) = \frac{4}{51} = 0.0784$

Litt mindre klart er kanskje at $P(E_2) = P(E_1) = \frac{4}{52}$, som er i overensstemmelse med 1).

[**Bevis:** Begivenheten E_2 kan splittes opp i 2 disjunkte deler:
$E_2 = (E_2 \cap E_1) \cup (E_2 \cap \overline{E}_1)$, slik at

$$P(E_2) = P(E_2 \cap E_1) + P(E_2 \cap \overline{E}_1) = P(E_1)P(E_2 \mid E_1) + P(\overline{E}_1)P(E_2 \mid \overline{E}_1) =$$
$$\frac{4}{52} \cdot \frac{3}{51} + \left(1 - \frac{4}{52}\right) \cdot \frac{4}{51} = \frac{4}{52 \cdot 51}(3 + 48) = \frac{4}{52}$$

Intuitivt kan vi forklare dette resultatet som følger: Tenk deg at du legger det første kortet du trekker ned på bordet med ryggen opp slik at du ikke vet om det er et ess eller ikke. Før du trekker ut kort nr. 2 så er det riktig at det bare er 51 kort igjen i kortstokken. Men like fullt er det 52 *mulige kort* du kan få siden du *ikke vet* hvilket kort som ligger på bordet. Av disse like sannsynlige *mulighetene* er det fire muligheter for å få ess. Det samme argumentet gjelder for alle de andre E_j-ene, slik at alle E_j-ene har samme sannsynlighet, $4/52$.]

Under forutsetningen om rent tilfeldig utvalg blir altså fordelingen for X gitt ved

$$P(X=x)=\frac{\binom{4}{x}\binom{48}{13-x}}{\binom{52}{13}}, \quad \text{for } x=0,1,2,\ldots,13$$

(Merk at $P(X=x)=0$ for $x>4$, siden $\binom{4}{x}=\frac{4\cdot 3\cdots 0\cdots(4-x+1)}{x!}=0$ for x > 4.)

Å beregne sannsynlighetene gjøres lett i Excel ved Excel-funksjonen, HYPGEOMDIST (sjekk selv):

Tabell 1. Fordeling hypergeom(52, 4, 13)

x	0	1	2	3	4	5	…
$P(X=x)$	0,304	0,439	0,213	0,041	0,0026	0	…

Den tilsvarende binomiske modellen for X er: $X \sim \text{bin}(13, p=4/52)$, som lett beregnes med BINOMDIST-funksjonen i Excel:

Tabell 2. Fordeling bin(13, 4/52)

x	0	1	2	3	4	5	..
$P(X=x)$	0,353	0,383	0,191	0,058	0,012	0,002	..

Merk at dette ville vært den korrekte fordelingen dersom vi hadde trukket hvert kort *med tilbakelegging*. Da legger vi altså hver gang det uttrukne kortet tilbake i kortstokken etter å registrert om det er ess eller ikke, og stokker kortene godt før neste trekning. I så fall blir

resultatene av enkelt-trekningene uavhengige, slik at begge de binomiske forutsetningene blir oppfylt – noe som nettopp impliserer at X er bin(13, 4/52) – fordelt.

Vi ser av tabell 1 og 2 at det i dette tilfellet er en relativ betydelig forskjell mellom de to fordelingene.

La oss nå øke populasjonsstørrelsen litt. Anta vi trekker 13 kort fra 4 kortstokker. Da blir

$$N = 52 \cdot 4 = 208, \quad M = 16, \quad n = 13$$

Vi får

$$P(X = x) = \frac{\binom{16}{x}\binom{192}{13-x}}{\binom{208}{13}}, \qquad \text{for } x = 0,1,2,\ldots,13$$

som utregnet (av Excel) gir

Tabell 3. Fordeling hypergeom(208, 16, 13)

x	0	1	2	3	4	5	…
$P(X = x)$	0,342	0,395	0,196	0,055	0,010	0,001	…

som ligger nærmere den binomiske fordelingen enn fordelingen i tabell 1.

Sluttmerknad. Hvis vi altså trekker et rent tilfeldig utvalg fra en større populasjon ($N > 10n$), kan vi således erstatte den hypergeometriske modellen, $X \sim \text{hypergeom}(N, M, n)$, med den enklere modellen, $X \sim \text{bin}(n,\ p = M/N)$. Dette er en stor fordel. I den binomiske modellen trenger vi for eksempel ikke å vite eksakt hvor stor populasjonen er (N) for å sette opp en fornuftig modell, så lenge populasjonen er stor nok. En annen fordel er at vi i situasjoner (som ofte forekommer) der X er antall suksesser i et utvalg trukket fra en populasjon, ikke trenger å begrunne realismen av de to binomiske forutsetningene (konstant suksess-sannsynlighet og uavhengige forsøk) for å kunne begrunne en binomisk modell for X. Det er nok å vise til at utvalget er rent tilfeldig og populasjonen er stor.