

Sandsynlighedsregning

2. forelæsning

Bo Friis Nielsen

Anvendt Matematik og Computer Science
Danmarks Tekniske Universitet
2800 Kgs. Lyngby – Danmark
Email: bfni@dtu.dk

Case: (motivation for Binomialfordelingen)

En fabrikant af plastik lyseholdere garanterer, at i en æske med 20 holdere vil højst to være itu. Man har gennem længere tid konstateret, at 5% af de producerede lyseholdere er defekte.

- Hvad er sandsynligheden for, at en æske overholder garantien?
- Hvad er sandsynligheden for, at alle holdere i en æske er intakte?

En forhandler køber et parti med 50 æsker. Han lover at erstatte alle holdere, der er itu.

- Hvad er det forventede antal holdere, han skal erstatte, hvis alle 50 æsker bliver solgt?

Vigtigste nye emner i 2.1, 2.2 og 2.5



- Binomialfordelingen p.81 : $P(k \text{ succeser}) = \binom{n}{k} p^k q^{n-k}$

- Normalfordelingsapproximationen p.99

$$P(\text{mellem } a \text{ og } b) = \Phi\left(\frac{b + \frac{1}{2} - n \cdot p}{\sqrt{n \cdot p \cdot q}}\right) - \Phi\left(\frac{a - \frac{1}{2} - n \cdot p}{\sqrt{n \cdot p \cdot q}}\right)$$

- Hypergeometrisk fordeling - stikprøve uden tilbagelægning - p.125

$$P(g \text{ gode og } b \text{ dårlige}) = \frac{\binom{G}{g} \binom{B}{b}}{\binom{N}{n}}$$

Det er bekvemt at generalisere problemet



- Antagelser:
 - ◊ Vi har n emner, med een af to egenskaber kaldet succes og fiasko
 - ◊ Hver har en sandsynlighed for succes på p
- Der er uafhængighed mellem de enkelte emner mht. egenskaberne succes/fiasko
- Vi har således n på hinanden følgende uafhængige Bernoulli eksperimenter

Bernoulli fordeling



- Det simplest mulige experiment
- To muligheder
 - ◊ Succes/fiasko
 - ◊ Mand/kvinde
 - ◊ Regn/ikke regn
- (I næste uge vil vi identificere den ene mulighed med 1, $X(\text{succes}) = 1$, den anden 0, $X(\text{fiasko}) = 0$)
 - ◊ $P(\text{succes}) = p$, $P(\text{fiasko}) = 1 - p$
- $X \sim be(p)$

For $n - 1$ intakte holdere finder vi



$$I_{n-1} = [(\cap_{i=1}^{n-1} H_i) \cap H_n^c] \cup [(\cap_{i=1}^{n-2} H_i) \cap H_{n-1}^c \cap H_n] \dots \cup [H_1^c \cap (\cap_{i=2}^n H_i)]$$

$$= \cup_{i=1}^n [H_i^c \cap (\cap_{j=1, j \neq i}^n H_j)]$$

$$P(I_{n-1}) = n \cdot (1 - p) \cdot p^{n-1}$$

- For I_i , generelt i , må vi tælle antallet af sekvenser med præcis i intakte. Antallet er

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}$$

vi kommer tilbage til udledningen af dette sidst i forelæsningen

- H_i : Hændelsen holder i intakt
- I_i : Hændelsen i intakte holdere
- Hvad er $P(I_i)$?
- Antag eventuelt, at holderne produceres sekventielt. Sandsynligheden for, at n holdere er intakte, er

$$P(n \text{ intakte}) = P(I_n) = P(H_1 \cap H_2 \cap \dots \cap H_n)$$

$$= P(\cap_{i=1}^n H_i) = \prod_{i=1}^n P(H_i) = p \cdot p \dots p = p^n$$

- Tilsvarende er sandsynligheden for, at alle n er itu:

$$P(n \text{ itu}) = P(I_0) = P(\cap_{i=1}^n H_i^c) = (1 - p) \dots (1 - p) = (1 - p)^n$$

Vi har udledt Binomial fordelingen



- $P(I_i) = \binom{n}{i} p^i (1 - p)^{n-i}$
- En fundamental byggesten/værktøj i ingeniørens værktøjskasse.
- Det forventede antal succeser ud af n (long term average)

$$E(X) = np$$

- Vi diskuterer dette begreb i dybden i næste uge - afsnit 3.2

Vi vil betragte et kast med 7 terninger.

Spørgsmål 1

Hvad er sandsynligheden for at få netop 4 seksere

- 1 $\frac{7! \cdot 125}{4!3! \cdot 6^7}$
- 2 $\frac{7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} \left(\frac{5}{6}\right)^3 \frac{1}{6^4}$
- 3 $\frac{4}{6}$
- 4 $\left(\frac{5}{6}\right)^3 \frac{1}{6^4}$
- 5 $\frac{4}{7}$
- 6 Ved ikke



- Fabrikanten har en stor detailhandelskæde mellem sine kunder
- Kæden indkøber 12.500.000 æsker
- kunden har konstateret 13.403.411 defekte holdere, som fabrikanten må erstatte ifølge kontrakten. Fabrikanten mistænker, at nogle af holderne er blevet ødelagt af kædens personale eller kunder.
- Hvem støtter I?
- Indledende undersøgelse:

$$\frac{13.403.411}{20 \cdot 12.500.000} = 0,0536$$

- Hvad skal man konkludere?

Tilbage til lyseholderfabrikanten



- Lad B_i betegne hændelsen, at i holdere er itu.

$$P(\text{Indenfor garantien}) = P(B_0) + P(B_1) + P(B_2) = P(I_{20}) + P(I_{19}) + P(I_{18})$$

- Sandsynlighederne $P(B_i)$ er givet ved Binomialfordelingen

$$\begin{aligned} &P(B_0) + P(B_1) + P(B_2) \\ &= 0,95^{20} + 20 \cdot 0,95^{19} \cdot 0,05 + \frac{20 \cdot 19}{2} \cdot 0,95^{18} \cdot 0,05^2 = 0,9245 \end{aligned}$$

- Det forventede antal, der skal erstattes fra en æske, er

$$np = 20 \cdot 0,05 = 1$$

- Vi forventer at måtte erstatte 50 holdere fra 50 æsker

Model

- Sandsynligheden $P(B_i)$ for hændelsen B_i , at i holderere er itu

$$P(B_i) = \binom{250.000.000}{i} 0,05^i 0,95^{250.000.000-i}$$

- Binomialfordelingen
- Vi ønsker at beregne

$$\sum_{i=13.403.411}^{250.000.000} P(B_i) = 1 - \sum_{i=0}^{13.403.410} P(B_i)$$

- En stærkt udfordrende beregning!

Normalfordeingen p.94

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$$



- Normalfordelingen er meget vigtig i sandsynlighedsregning og statistik
- Blandt andet på grund af den centrale grænseværdisætning (CGS - eng: Central Limit Theorem - CLT), der siger, at gennemsnittet af et stort antal uafhængige målinger er meget godt beskrevet ved normalfordelingen. Dette gælder under meget generelle antagelser.

Lad H_i betegne hændelsen: Antallet af krone i 400 kast med en retfærdig mønt er i .



Spørgsmål 2

Beregn sandsynligheden for $P(\cup_{i=190}^{210} H_i)$ - (eventuelt approksimativt).

- $\frac{1}{2}$
- $\Phi(1,05) - \Phi(-1,05)$
- $\Phi(1,75) - \Phi(-1,75)$
- $\sum_{i=185}^{215} \binom{400}{i} \frac{1}{2^{400}}$
- $\Phi(1,50) - \Phi(-1,50)$
- Ved ikke
hvor Φ betegner fordelingsfunktionen for en standardiseret normalfordelt variabel.

Normalapproximationen p.99



- Antal succeser beskrives ved $B(n, p)$

$$\begin{aligned} P(a \leq \text{antal succeser} \leq b) &= \sum_{i=a}^b \binom{n}{i} p^i (1-p)^{n-i} \\ &= \Phi\left(\frac{b - np + \frac{1}{2}}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - np - \frac{1}{2}}{\sqrt{np(1-p)}}\right) \end{aligned}$$

- Brugen af $a - \frac{1}{2}$ og $b + \frac{1}{2}$ snarere end a, b kaldes en kontinuitetskorrektion.
- Den er vigtig for små værdier af $\sqrt{np(1-p)}$
- Approximationen kan forfines yderligere: side 104

Normalapproximationen for lyseholderne

- Vi ønsker at beregne $P(\text{mindre end } 13.403.411 \text{ defekte holdere})$, som ved brug af normalapproximationen findes til

$$\begin{aligned} \Phi\left(\frac{13.403.410 + \frac{1}{2} - 250.000.000 \cdot 0,05}{\sqrt{250.000.000 \cdot 0,05(1 - 0,05)}}\right) \\ = \Phi\left(\frac{903.410,5}{3.446.01}\right) = \Phi(262,16) \approx 1 \end{aligned}$$

- Den sandsynlighed vi er interesseret i er $P(\text{mere end } 13.403.410)$, der er ≈ 0 .
- Enten har fabrikanten været meget uheldig, eller fejlraten er højere end 5%, eller nogle af lyseholderne er blevet beskadiget efter fabrikationen.

Hvad er rimelige værdier for antallet af defekte holdere?



- Vi vil bestemme c , således at

$$P(n \cdot p - c \leq \text{Antal defekte holdere} \leq n \cdot p + c) = 0,95$$

- Ved brug af normalapproximationen får vi

$$\begin{aligned} &P(n \cdot p - c \leq \text{Antal defekte holdere} \leq n \cdot p + c) \\ &= \Phi\left(\frac{(n \cdot p + c) + \frac{1}{2} - n \cdot p}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{(n \cdot p - c) - \frac{1}{2} - n \cdot p}{\sqrt{np(1-p)}}\right) \end{aligned}$$

Med

$$2\Phi\left(\frac{c}{\sqrt{np(1-p)}}\right) - 1 = 0,95$$



får vi

$$\Phi\left(\frac{c}{\sqrt{np(1-p)}}\right) = \frac{1,95}{2} = 0,975$$

Vi har

$$\Phi^{-1}(0,975) = 1,96$$

Således at

$$\frac{c}{\sqrt{np(1-p)}} = 1,96 \quad c = 1,96 \cdot \sqrt{np(1-p)} = 6754,2$$

- Der vil altså med 95 % sandsynlighed være mellem 12.493.246 og 12.506.754 defekte holdere.
- **Et overraskende snævert interval**

- For store n er dette approksimativt



$$\begin{aligned} &\Phi\left(\frac{c}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{-c}{\sqrt{np(1-p)}}\right) \\ &= \Phi\left(\frac{c}{\sqrt{np(1-p)}}\right) - \left(1 - \Phi\left(\frac{c}{\sqrt{np(1-p)}}\right)\right) \\ &= 2\Phi\left(\frac{c}{\sqrt{np(1-p)}}\right) - 1 \end{aligned}$$

Kvadratrodsloven p.100



- For store n ; i n uafhængige forsøg med sandsynlighed p for succes i det enkelte forsøg
 - ◇ vil antallet af succeser med høj sandsynlighed, ligge i et relativt snævert interval centreret om np , med et moderat multiplum af \sqrt{n} på en numerisk skala.
- *Andelen* af succeser vil, med høj sandsynlighed, ligge i et smalt interval centreret om p , med en bredde, der er et moderat multiplum af $\frac{1}{\sqrt{n}}$.

De store tals lov p.101



- Hvis n er stor, vil andelen af succeser i n , uafhængige forsøg, med overvældende sandsynlighed, være meget tæt på p , sandsynligheden for succes i det enkelte forsøg. Mere formelt:
 - ◇ for uafhængige forsøg, med sandsynlighed p for succes i hvert forsøg, for hver $\epsilon > 0$, uanset størrelse, for $n \rightarrow \infty$

$$P(\text{Andelen af succeser i de } n \text{ forsøg afviger mindre end } \epsilon \text{ fra } p) \rightarrow 1$$



By A har en befolkning på 4 millioner, og by B har en befolkning på 6 millioner. Begge byer har den same andel af kvinder. Man tager en tilfældig stikprøve af personer (med tilbagelægning), fra *hver* by for at estimere denne andel. Følgende tre metoder til at bestemme stikprøvestørrelserne overvejes

- ◇ I: En 0,01% stikprøve fra hver by
- ◇ II: En stikprøve på 400 fra hver by.
- ◇ III: En 0,1% stikprøve fra by A, og en 0,075% stikprøve fra by B.

Spørgsmål 3



- Angiv, for hvilke metoder estimatet p_A for by A og estimatet p_B for by B har samme præcision.

- Ingen af metoderne giver samme præcision for A og B
- Kun metode I giver samme præcision for A og B
- Kun metode II giver samme præcision for A og B
- Metode I og II giver samme præcision for A og B
- Metode II og III giver samme præcision for A og B
- Ved ikke



- En 0,01% stikprøve fra hver by
 - ◇ Stikprøvestørrelserne er 400 og 600; præcisionen af estimerterne er proportional med $\frac{1}{\sqrt{400}}$ respektivt $\frac{1}{\sqrt{600}}$
- En stikprøve på 400 fra hver by.
 - ◇ Vi får lige gode estimerter
- En 0,1% stikprøve fra by A, og en 0,075% stikprøve fra by B.
 - ◇ Stikprøvestørrelserne er 4.000 og 4.500, stikprøven fra B er en lille smule bedre.

Kombinatorik (counting App. 1.)

- Hvis vi har N forskellige elementer, på hvor mange måder kan vi da udtage en stikprøve på $n \leq N$ uden tilbagelægning?
 - ◊ Det første element kan udtages på N forskellige måder.
 - ◊ Det andet element skal udtages ud af en population på $N - 1$ individer, hvilket således kan ske på $N - 1$ måder.
 - ◊ Alt i alt kan vi udtage en stikprøve på n elementer ud af en population på N , på

$$(N)_n = N(N - 1) \dots (N - n + 1) = \frac{N!}{(N - n)!}$$

måder.

- Hvor $(N)_n$ betegner antallet af ordnede stikprøver af størrelse n ud af N elementer

Stikprøvetagning med tilbagelægning



- n elementer udtages ud af en population med N elementer, hvor alle elementer har samme sandsynlighed for at blive udtrukket og således, at det samme emne kan udtages hver gang

Stikprøvetagning uden tilbagelægning

- n elementer udtages ud af en population med N elementer, hvor alle elementer har samme sandsynlighed for at blive udtrukket og således, at et emne kun kan udtrækkes en gang

Uden ordning af elementerne

- Vi indfører symbolet $\binom{N}{n}$ til at betegne antallet af måder, hvorpå vi kan udtage en stikprøve af størrelse n ud af N uden ordning.
- Vi kan få en ordnet stikprøve ved at ordne en uordnet,
- Der er $n!$ forskellige måder, hvorpå vi kan ordne n elementer, så

$$(N)_n = \frac{N!}{(N - n)!} = \binom{N}{n} n!$$

- Vi har således $\binom{N}{n} = \frac{N!}{n!(N - n)!}$

Den hypergeometriske fordeling



- Vi betragter en population, med N individer, hvor hvert individ er karakteriseret ved besiddelse eller mangel af en egenskab - succes/fiasco - modellen.
- Når vi udtager n elementer uden tilbagelægning ændrer den relative forekomst af individer med egenskaben sig mellem hver trækning.
- Vi kan bestemme sandsynligheden for ud af en stikprøve på n at få netop g med egenskaben.
- Sandsynligheden bestemmes som antallet af måder, hvormed vi kan udtage en stikprøve med netop g "positive" i forhold til det totale antal af mulige stikprøver.

Hypergeometrisk fordeling fortsat



- Antallet af måder hvormed vi kan udtage g positive elementer på er $\binom{G}{g}$
- Antallet af måder, hvormed vi kan udtage $b = n - g$ negative elementer på er $\binom{N - G}{n - g}$
- Antallet af måder, hvormed vi kan udtage n elementer ud af N på er $\binom{N}{n}$



- så sandsynligheden for hændelsen A_g at få netop g positive elementer ud af n er

$$P(A_g) = \frac{\binom{G}{g} \binom{N - G}{n - g}}{\binom{N}{n}}$$

Eksempel på hypergeometrisk fordeling



- Et firma, der har været belastet af et dårligt arbejdsklima, har foretaget en anonym tilfredshedsundersøgelse blandt medarbejderne.
- Denne undersøgelse har vist, at der i en afdeling med 19 medarbejdere er 7, der må forventes indledningsvist at modarbejde ledelsesmæssige tiltag til forbedring af arbejdsklimaet.
- Et konsulentfirma har anbefalet at nedsætte en konstruktiv arbejdsgruppe på 3 personer, der skal komme med forslag.
- Det forventes, at gruppen kan præstere nyttige resultater, hvis der er mindst to konstruktive medarbejdere i arbejdsgruppen.

- Vurder mulighederne for, at arbejdsgruppen får succes
- Nedsættelse af arbejdsgruppen kan betragtes som stikprøvetagning uden tilbagelægning.
- Vi definerer hændelserne K_i som hændelsen, at der er netop i konstruktive personer i arbejdsgruppen

$$\begin{aligned} P(\text{succes}) &= P(K_2 \cup K_3) = P(K_2) + P(K_3) \\ &= \frac{\binom{12}{2} \binom{7}{1}}{\binom{19}{3}} + \frac{\binom{12}{3} \binom{7}{0}}{\binom{19}{3}} = \frac{\frac{12 \cdot 11}{2} \cdot 7 + \frac{12 \cdot 11 \cdot 10}{6}}{\frac{19 \cdot 18 \cdot 17}{3 \cdot 2}} = \frac{682}{969} \end{aligned}$$

Afsnit 2.1, 2.2 og 2.5



- Binomialfordelingen p.81 : $P(k \text{ succeser}) = \binom{n}{k} p^k q^{n-k}$

- Normalfordelingsapproximationen p.99

$$P(\text{mellem } a \text{ og } b) = \Phi\left(\frac{b + \frac{1}{2} - n \cdot p}{\sqrt{n \cdot p \cdot q}}\right) - \Phi\left(\frac{a - \frac{1}{2} - n \cdot p}{\sqrt{n \cdot p \cdot q}}\right)$$

- Hypergeometrisk fordeling - stikprøve uden tilbagelægning - p.125

$$P(g \text{ gode og } b \text{ dårlige}) = \frac{\binom{G}{g} \binom{B}{b}}{\binom{N}{n}}$$