

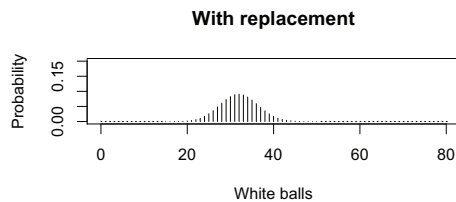
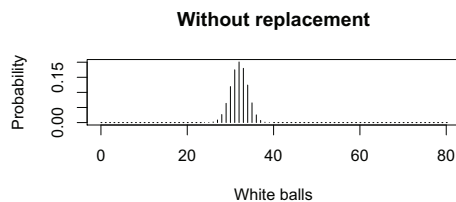
Sandsynlighedsregning

4. forelæsning

Bo Friis Nielsen

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

Standardafvigelse/varsians



Den hypergeometriske fordeling har mindre variation end binomialfordelingen.

Dagens emner: Afsnit 3.3 og 3.4

- Varians/standardafvigelse

$$\text{Var}(X) = E[(X - E(X))^2] = \sum_{\text{alle } x} (x - E(X))^2 P(X = x)$$

- Normalfordelingsapproximation/den Centrale Grænseværdisætning

$$P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a)$$

- Markovs og Chebychevs uligheder for ekstreme udfald

$$P(X \geq a) \leq \frac{E(X)}{a} \quad P(|X - E(X)| \geq kSD(X)) \leq \frac{1}{k^2}$$

- Et udpluk af diskrete fordelinger

$$P(T = i) = (1 - p)^{i-1} p \quad P(Y_r = i) = \binom{i+r-1}{r-1} p^r (1-p)^i$$

- Indikatorfunktioner $I_A = \begin{cases} 1 & \text{hvis } A \text{ indtræffer} \\ 0 & \text{hvis } A^c \text{ indtræffer} \end{cases}$

Definition af varsians

$$\text{Var}(X) = E[(X - E(X))^2] = \sum_{\text{alle } x} (x - E(X))^2 P(X = x)$$

Beregningsformel for varsians (vigtig) p.186

$$\text{Var}(X) = E(X^2) - (E(X))^2$$

“Shift and scale”

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

Sum af uafhængige variable

Hvis X_1, \dots, X_n er uafhængige, så er

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$$



Varians af et tal trukket fra en liste



Givet en liste af tal

$$(x_1, x_2, \dots, x_n)$$

Lad X være en stokastisk variabel: Et tilfældigt element af listen.

$$E(X) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

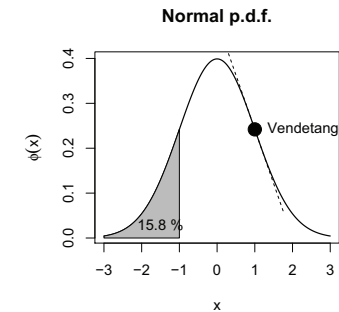
Standardafvigelse



$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

“Shift and scale”

$$\text{SD}(aX + b) = |a|\text{SD}(X)$$



Bestem variansen i binomialfordelingen (n, p)



Først, skriv $X \sim \text{Bin}(n, p)$ som

$$X = X_1 + \dots + X_n \quad \text{hvor} \quad X_i \sim \text{Bernoulli}(p)$$

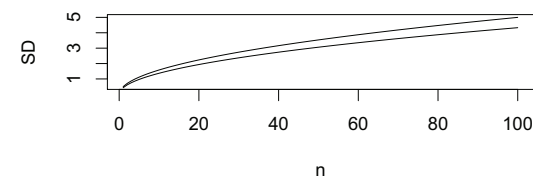
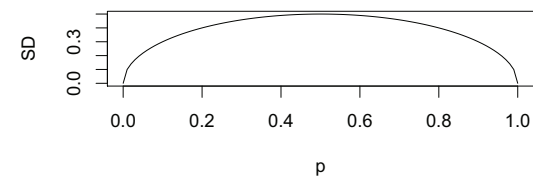
Dernæst, variansen i en Bernoullifordeling (p):

$$E(X_i) = p, \quad E(X_i^2) = p, \quad \text{Var}(X_i) = p - p^2 = p(1 - p)$$

Kombinér:

$$\text{Var}(X) = np(1 - p)$$

Spredningen i binomialfordelingen (n, p)



Lad X være antallet af dage i en måned valgt tilfældigt blandt årets tolv måneder i et ikke-skudår. Det anføres, at $E(X) = \frac{365}{12}$.

Spørgsmål 1

Man finder $SD(X)$ til

- 1 0,71
- 2 0,76
- 3 0,81
- 4 0,86
- 5 0,91
- 6 Ved ikke

Kvadratrodsloven



Lad X_i være I.I.D. (Independent and Identically Distributed).

Definér summen $S_n = X_1 + \dots + X_n$

Definér gennemsnittet $\bar{X}_n = \frac{1}{n}S_n$.

$$E(S_n) = nE(X_i) \quad , \quad SD(S_n) = \sqrt{n}SD(X_i)$$

$$E(\bar{X}_n) = E(X_n) \quad , \quad SD(\bar{X}_n) = \frac{1}{\sqrt{n}}SD(X_n)$$

Store tals svage lov



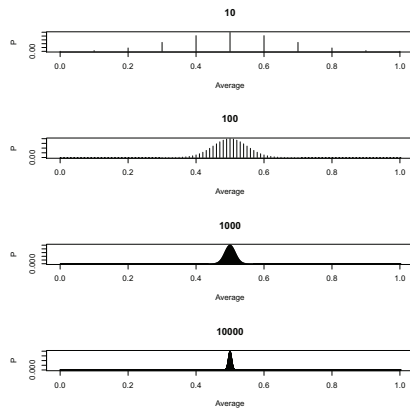
Lad X_i være I.I.D. med middelværdi μ og spredning σ .

$$\forall \epsilon > 0 :$$

$$n \rightarrow \infty \Rightarrow$$

$$P(|\bar{X}_n - \mu| < \epsilon) \rightarrow 1$$

(\bar{X}_n konvergerer mod μ i sandsynlighed)



Den centrale grænseværdisætning p.196



Lad X_1, \dots, X_n være I.I.D. med middelværdi μ og varians σ^2 .

Definér $S_n = X_1 + \dots + X_n$.

For store n gælder approksimativt

$$P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a)$$

Betydning



- CGS er et centralt resultat i sandsynlighedsregningen og statistikken (og findes i mange varianter)
- Budskabet er, at stokastiske variable, der er dannet gennem en sum af mange mindre bidrag, med god tilnærmelse kan beskrives ved en normalfordeling, som vi strengt taget først stifter bekendtskab med i næste uge
- Vi har allerede set denne sætning i anvendelse for binomialfordelingen (og Poissonfordelingen)

Om en stokastisk variabel X , der beskriver et rumindhold, oplyses det, at den har middelværdi $E(X) = 10$.



Spørgsmål 2

Den mindste øvre grænse for $P(X \geq 20)$ findes til

- $\frac{1}{4}$
- $\frac{1}{2}$
- $\frac{3}{4}$
- $1 - \Phi\left(\frac{20-10}{\sqrt{20}}\right)$
- En sådan grænse kan ikke bestemmes ud fra det oplyste
- Ved ikke

Hvor Φ som sædvanlig betegner fordelingsfunktionen for en standardiseret normalfordelt variabel.

Markovs ulighed p.174



- For ikke negative stokastiske variable, kan vi angive en øvre grænse for sandsynligheder ved brug af middelværdien.

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Chebychevs ulighed p.191

- Hvis vi også kender variansen kan vi som regel skærpe denne grænse

$$P(|X - E(X)| \geq kSD(X)) \leq \frac{1}{k^2}$$

Et udvalg af diskrete fordelinger



- En række grundlæggende mekanismer
- Se pp.475 “Distribution summaries”

Spørgsmål 3



Hvis en række tal c_i udgør en sandsynlighedsfordeling, hvad må der så gælde om c_i ?

- 1 Der er ingen specifikke krav
- 2 $\sum_i c_i = 1$
- 3 c_i kan bestemmes fra en formel som $c_i = \binom{n}{i} p^i (1-p)^{n-i}$
- 4 $c_i \geq 0$
- 5 Ingen af de ovenstående er tilstrækkelige
- 6 Ved ikke

T_r Antal forsøg indtil r te succes



Negativ binomialfordeling

- Y_r Antal fiaskoer til den r 'te succes i en sekvens af Bernoulliforsøg
- Hvor mange kameraer skal vi kassere før vi får r , der passerer kvalitetskontrollen

$$T_r = Y_r + r$$

- (Har i øvrigt mange andre fortolkninger)

Den geometriske fordeling



T : antal Bernoulli forsøg indtil første succes

$$P(T = i) = (1-p)^{i-1} p$$

$$P(T > i) = (1-p)^i$$

Middelværdi

$$E(T) = \sum_{i=0}^{\infty} P(T > i) = \frac{1}{p}$$

Varians

$$\text{Var}(T) = \frac{1-p}{p^2}, \quad \text{SD}(T) = \frac{\sqrt{1-p}}{p}$$

Spørgsmål 4



Hvad er middelværdi og varians for en $NB(r, p)$ fordeling?

- 1 $E(Y_r) = \frac{r(1-p)}{p}, \text{Var}(Y_r) = \frac{r(1-p)}{p^2}$
- 2 $E(Y_r) = \frac{r(1-p)}{p}, \text{Var}(Y_r) = \frac{(1-p)}{p^2}$
- 3 $E(Y_r) = \frac{1}{p}, \text{Var}(Y_r) = \frac{r(1-p)}{p^2}$
- 4 $E(Y_r) = \frac{1}{p}, \text{Var}(Y_r) = \frac{(1-p)}{p^2}$
- 5 $E(Y_r) = \frac{r}{p}, \text{Var}(Y_r) = \frac{r}{p^2}$
- 6 Ved ikke

Negativ binomialfordeling udledning



Antal fiaskoer før r 'te succes.

En sekvens af Bernoulli forsøg: fffsffsf

Vi skal placere $r - 1$ succeser på $i + r - 1$ forsøg før vores "endelige" succes.

Derfor

$$\begin{aligned} P(Y_r = i) &= \binom{i+r-1}{r-1} p^{r-1} (1-p)^i p \\ &= \binom{i+r-1}{r-1} p^r (1-p)^i \end{aligned}$$

Indikatorfunktionen (p.155), p.168



$$I_A = \begin{cases} 1 & \text{hvis } A \text{ indtræffer} \\ 0 & \text{hvis } A^c \text{ indtræffer} \end{cases}$$

Spørgsmål 5

Hvad er $E(I_A)$

- 1 0,5
- 2 1
- 3 A
- 4 $P(A)$
- 5 Kan man ikke sige generelt
- 6 Ved ikke

Middelværdi og varians



$$Y_r \sim NB(r, p) \quad Y_r = T_r - r \quad T_r = W_1 + \dots + W_r$$

hvor W_i er ventetiden fra den $i - 1$ te til den i te succes.

W_i er IID; geometrisk (p) fordelt: Lad X_i være antallet af fiaskoer mellem succes $i - 1$ og i .

$$E(W_i) = \frac{1}{p} \quad \text{Var}(W_i) = \frac{1-p}{p^2}$$

Vi har da

$$Y_r = \sum_{i=1}^r X_i = \sum_{i=1}^r (W_i - 1) = \sum_{i=1}^r W_i - r$$

Derfor:

$$E(Y_r) = \frac{r}{p} - r = r \frac{1-p}{p} \quad \text{Var}(Y_r) = r \frac{1-p}{p^2}$$

Hvornår kan man bruge indikatorfunktioner?



Man kan bruge indikatorfunktionen når vi kan skrive en stokastisk variabel X som

$$X = I_{A_1} + I_{A_2} + \dots + I_{A_k}$$

- dvs. X tæller, hvor mange af hændelserne A_1, \dots, A_k , der er indtruffet.

$$E(X) = P(A_1) + P(A_2) + \dots + P(A_k)$$

bruges når $P(A_i)$ kan bestemmes (let), men fordelingen af X ikke kan. (F.eks. opgave 3.2.14)

Opgave 3.3.8



Lad A_1 , A_2 og A_3 være hændelser med sandsynlighederne

$$P(A_1) = \frac{1}{5}, \quad P(A_2) = \frac{1}{4}, \quad P(A_3) = \frac{1}{3}$$

Lad N betegne antallet af disse hændelser, der indtræffer.

Opskriv N udtrykt ved indikatorvariable

- Variablen I_B antager værdien 1, hvis B indtræffer, ellers 0.

$$N = I_{A_1} + I_{A_2} + I_{A_3}$$

Beregn $\text{Var}(N)$ når



- A_1 , A_2 og A_3 er gensidigt udelukkende
- A_1 , A_2 og A_3 er uafhængige
- $A_1 \subset A_2 \subset A_3$

Find $E(N)$



$$N = I_{A_1} + I_{A_2} + I_{A_3}$$

Vi tager forventning på begge sider

$$E(N) = E(I_{A_1} + I_{A_2} + I_{A_3})$$

Dermed

$$E(N) = E(I_{A_1}) + E(I_{A_2}) + E(I_{A_3}) \quad \text{Hvorfor?}$$

$$E(N) = P(A_1) + P(A_2) + P(A_3) = \frac{47}{60}$$

$\text{Var}(N)$ når A_i er gensidigt udelukkende



I dette tilfælde har vi

$$P(N > 1) = 0$$

$$P(N = 1) = P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) = E(N)$$

Eksperimentet er altså et Bernoulliekperiment og vi har

$$\text{Var}(N) = p(1 - p) = \frac{47}{60} \frac{13}{60} = 0,170 = 0,412^2$$

Var(N) når A_1 , A_2 og A_3 er uafhængige



$$\text{Var}(N) = \text{Var}(I_{A_1} + I_{A_2} + I_{A_3}) = \text{Var}(I_{A_1}) + \text{Var}(I_{A_2}) + \text{Var}(I_{A_3})$$

For hver A_i har vi

$$\text{Var}(I_{A_i}) = P(A_i) \cdot P(A_i^c)$$

Altså

$$\text{Var}(I_{A_1}) = \frac{1}{5} \cdot \frac{4}{5} = \frac{4}{25} \quad \text{Var}(I_{A_2}) = \frac{3}{16} \quad \text{Var}(I_{A_3}) = \frac{2}{9}$$

Så

$$\text{Var}(N) = \frac{2051}{3600} \approx 0,570 \approx 0,755^2$$

Beregn $\text{Var}(N)$ når



A_1 , A_2 og A_3 er gensidigt udelukkende $\text{Var}(N) = 0,412^2$

A_1 , A_2 og A_3 er uafhængige $\text{Var}(N) = 0,755^2$

$A_1 \subset A_2 \subset A_3$ $\text{Var}(N) = 1,17^2$

Kan vi forklare/forstå resultatet?

Beregn $\text{Var}(N)$ når $A_1 \subset A_2 \subset A_3$



$$P(N = 3) = P(A_1) = \frac{1}{5}$$

Vi har ved brug af differensreglen p.22

$$P(N = 2) = P(A_2 \setminus A_1) = P(A_2) - P(A_1) = \frac{1}{20}$$

$$P(N = 1) = P(A_3 \setminus A_2) = P(A_3) - P(A_2) = \frac{1}{12}$$

Dermed:

$$E(N^2) = \frac{9}{5} + \frac{4}{20} + \frac{1}{12} = \frac{119}{60}$$

$$\text{Var}(N) = E(N^2) - (E(N))^2 = \frac{4931}{3600} = 1,370 = 1,170^2$$

Tail sum formula



Hvis X kan antage værdier $0, 1, \dots$

$$E(X) = \sum_{i=1}^{\infty} P(X \geq i)$$

(En tilsvarende formel gælder for kontinuerte variable; kap. 4)

Bevis:

$$X = \sum_{i=1}^X 1 = \sum_{i=1}^{\infty} I_{X \geq i}$$

$$E(X) = E\left(\sum_{i=1}^{\infty} I_{X \geq i}\right) = \sum_{i=1}^{\infty} E(I_{X \geq i}) = \sum_{i=1}^{\infty} P(X \geq i)$$

Et par nyttige sumformler



$$\sum_{i=0}^N a^i = \frac{1 - a^{N+1}}{1 - a}, \quad |a| \neq 1$$

$$\sum_{i=0}^{\infty} a^i = \frac{1}{1 - a}, \quad |a| < 1$$

$$\sum_{i=0}^{\infty} \frac{a^i}{i!} = e^a$$

- For $a_i \geq 0$ og $b_i \geq 0$ har vi:

$$\left(\sum_{i=0}^{\infty} a_i \right) \left(\sum_{i=0}^{\infty} b_i \right) = \sum_{i=0}^{\infty} \left(\sum_{k=0}^i a_k b_{i-k} \right) = \sum_{i=0}^{\infty} c_i \quad c_i = \sum_{k=0}^i a_k b_{i-k}$$

Nye begreber i afsnit 3.3 og 3.4



- Varians og standardafvigelse
- Markovs og Chebychevs uligheder
- Indikatorfunktion
- Den centrale grænseværdisætning (3.3)
- Geometrisk fordeling
- Negativ binomial fordeling

Sammenhæng mellem fordelinger



- $X \sim \text{Bin}(n_1, p)$ og $Y \sim \text{Bin}(n_2, p)$: $X + Y \sim \text{Bin}(n_1 + n_2, p)$
- $X \sim \text{Geom}(p)$: $X - 1 \sim \text{NB}(1, p)$.
- $X \sim \text{NB}(r_1, p)$ og $Y \sim \text{NB}(r_2, p)$: $X + Y \sim \text{NB}(r_1 + r_2, p)$
- $\text{HyperGeom}(n, N, G) \rightarrow \text{Bin}(n, p)$ når $N, G \rightarrow \infty$ mens $G/N \rightarrow p$.
- Summer kan generelt approksimeres med normalfordelingen

Afsnit 3.3 og 3.4

- Varians/standardafvigelse
- Normalfordelingsapproximation/den Centrale Grænseværdisætning

$$\text{Var}(X) = E[(X - E(X))^2] = \sum_{\text{alle } x} (x - E(X))^2 P(X = x)$$

$$P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a)$$

- Markovs og Chebychevs uligheder for ekstreme udfald

$$P(X \geq a) \leq \frac{E(X)}{a} \quad P(|X - E(X)| \geq kSD(X)) \leq \frac{1}{k^2}$$

- Et udpluk af diskrete fordelinger

$$P(T = i) = (1-p)^{i-1}p \quad P(T_r = i) = \binom{x+r-1}{r-1} p^{r-1}(1-p)^i$$

- Indikatorfunktioner $I_A = \begin{cases} 1 & \text{hvis } A \text{ indtræffer} \\ 0 & \text{hvis } A^c \text{ indtræffer} \end{cases}$