

Binomialfordelingen:

april 09 GJ

Spm A-14: Sandsynlighedsregning og statistik.

Efter en kort introduktion af grundlæggende begreber i sandsynlighedsregning og statistik skal du skal introducere binomialfordelingsmodellen og udvikle formlen for hvordan binomialsandsynligheder udregnes.

Specielt Notatet om binomialfordelingen med henvisninger

Sandsynlighedsregning og statistik

Bemærkning:

1 – sandsynlighedsregning	2 - Deskriptiv statistik:
I sandsynlighedsregning opstiller man matematiske modeller for hvordan et tænkt eksperiment vil falde ud. Det er altså en teoretisk tilgang, der bygger på viden om eksperimentet.	I deskriptiv statistik observerer man hvordan et konkret eksperiment faktisk forløb. Observationssætter repræsenterer hvad vi så, og ikke hvad vi <i>tænkte</i> .
Udfaldsrum – det vi kan få	Observationssæt – det vi faktisk observerede
Sandsynligheder for udfald	Frekvenser
Vi skelner mellem et endeligt udfaldsrum som i binomialfordelingen , hvor vi har n forskellige muligheder og så den anden type som fx normalfordelingen , hvor der er uendelig mange muligheder (fx elevernes højde i en en klasse) • Udfald, sandsynlighed, hændelser, • regneregler for hændelser • uafhængige hændelser	Hvis der er mange gentagelser af observationer vil man regne på observationssættet som det er (fx antal 6'ere ved 4 kast med en terneing). Hvis der er få eller ingen gentagelser (fx skattepligtige indkomst for gymnasie lærerne på FG) vil man gruppere observationerne i intervaller. For ikke-grupperede observationssæt: • observationshyppighed, observationsfrekvens og pindediagram • kvartilsæt, median • middeltal • boxplot For grupperede observationer: • interval-hyppighed, interval-frekvens, histogram • kumuleret frekvens, sumkurve, kvartilsæt, median

Binomialfordeling

Bemærkning: Dette er helt centralt i spørgsmålet. I bogen forklares hvordan man kommer frem til binomialsandsynlighederne mens formlen for binomialkvotienten ikke forklares. Nedenfor er givet en folklaring/et bevis på begge dele.

Tekst: Du kan godt nøjes med det der står nedenfor (Du kan finde en anden måde at forklare nogle af de samme ting i bogen II 158 II-arbejdsbog 130-133. Du behøver ikke benytte bogen)

Definition af binomialfordeling

Udføres et *basiseksperiment* med sandsynligheden p for succes og sandsynligheden $1-p$ for fiasko. Lad X tælle antal succes'er, når vi udfører basiseksperimentet n gange. Vi siger at X er binomialfordelt med antalsparamenter n og sandsynlighedsparemeter p . Vi skriver $X \sim bi(n,p)$.

Definition af binomialkvotient

Antallet af r -delmængder af en n -mængde kaldes $K(n,r)$

Formel til bestemmelse af binomialsandsynlighed

Sandsynligheden for at få netop i succeser når vi udfører det samme basiseksperiment n gange kan bestemmes ved følgende formel:

$$P(i \text{ succes'er}) = P(X=i) = K(n,i) \cdot p^i \cdot (1-p)^{n-i}$$

hvor p angiver sandsynligheden for succes i basiseksperimentet, der er det eksperiment der gentages i alt n gange.

Formel til bestemmelse af binomialkvotient

Antallet af måder vi kan udvælge en r -delmængde af en n -mængde beregnes ved følgende formel:

$$K(n,r) = \frac{n \cdot (n-1) \cdot (n-2) \dots \text{I alt } r \text{ faktorer}}{r \cdot (r-1) \cdot (r-2) \dots 2 \cdot 1 \text{I alt } r \text{ faktorer}} = \frac{n!}{r! \cdot (n-r)!}$$

$K(n,r)$ er altså en brøk hvor der for oven står de r første tal i rækken $n \cdot (n-1) \cdot (n-2) \dots$ og for neden står de r sidste tal ganget sammen.

Lad os "bevise" formlen for binomialsandsynlighed ud fra et par konkrete problem.

1 - Eksperiment 1 – antal 6'ere i 3 kast med terning

Vi kaster en terning 3 gange og tæller antal 6'ere. Basiseksperimentet er at kaste en terning en gang og se om det bliver en 6'er. Sandsynlighedsparameteren er altså $p=1/6$ og antalsparameteren er $n=3$

Lad os beskrive ved SSF hvis vi får succes (dvs 6'er i 1. og 2. kast, mens vi får fiasko (dvs ikke-6'er) i 3. kast.

$P(SSF) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6}$ eller $p^2 \cdot (1-p)^1$ hvis p betegner sandsynligheden for succes i basiseksperimentet.

Vi kan udregne sandsynligheden for to succeser ved

$$P(X=2) = P(2 \text{ succeser}) = P(SSF) + P(SFS) + P(FSS) = 3 \cdot p^2 \cdot (1-p)^1$$

Bemærk at tallet 3 angiver hvor mange måder vi kan fordele vores succes'er og fiasko'er på. Vi skal i dette tilfælde vælge 2 ud af de tre pladser {1. plads, 2. plads, 3. plads}. Vi skal altså vælge en 2-delmængde ud af en 3-mængde, og det kan vi gøre på $K(3,2)$ måder. Det er i dette tilfælde klart, at det kan gøres på 3 måder. Hvis vi skulle beregne det ud fra formlen ville det give

$$K(3,2) = \frac{3!}{2! \cdot (3-2)!} = \frac{3 \cdot 2 \cdot 1}{(2 \cdot 1) \cdot (1)} = 3$$

Eller

$$K(3,2) = \frac{3 \cdot 2}{2 \cdot 1} = 3$$

Skal vi lave en oversigt over alle de muligheder, så kan vi få 0, 1, 2 eller 3 succeser i dette eksperiment. Sandsynlighederne er

$$P(X=0) = P(0 \text{ succeser}) = P(FFF) = (1-p)^3 = 1 \cdot p^0 \cdot (1-p)^3$$

$$P(X=1) = P(1 \text{ succeser}) = P(SFF) + P(FSF) + P(FFS) = 3 \cdot p^1 \cdot (1-p)^2$$

$$P(X=2) = P(2 \text{ succeser}) = P(SSF) + P(SFS) + P(FSS) = 3 \cdot p^2 \cdot (1-p)^1$$

$$P(X=3) = P(3 \text{ succeser}) = P(SSS) = p^3 = 1 \cdot p^3 \cdot (1-p)^0$$

2 - Eksperiment 2 – antal 6'ere i 5 kast med terning

Vi starter med at bestemme sandsynligheden for at få to succes'er i dette eksperiment. Det kunne være ved at de to første var 6'ere og de sidste 3 kast viste ikke-6'ere, eller det kunne være det første og det 5. kast der var 6'ere. Opskriver vi samtlige muligheder får vi

$$P(X=2) = P(2 \text{ succes'er}) = P(SSFFF) + P(SFSFF) + P(SFFSF) + P(SFFFS) + P(FSSFF) + P(FSFSF) + P(FSFFS) + P(FFSSF) + P(FFSFS) + P(FFFSS) = 10 \cdot p^2 \cdot (1-p)^3$$

Igen ser vi at tallet 10 angiver hvor mange måder vi kan fordele vores succes'er på. Vi skal i dette tilfælde vælge 2 ud af de 5 pladser {1. plads, 2. plads, 3. plads, 4. plads, 5. plads}. Vi skal altså vælge en 2-delmængde ud af en 5-mængde, og det kan vi gøre på $K(5,2)$ måder.

Vi har set, at dette åbenbart kan gøres på 10 måder. Hvis vi skulle beregne det ud fra formlen ville det give

$$K(5,2) = \frac{5 \cdot 4}{2 \cdot 1} = 10$$

3 - En generel argumentation for formlen

Vi beregner sandsynligheden for i succeser i en binomialfordeling med $X \sim \text{bi}(n,p)$ ved

$$P(i \text{ succes'er}) = P(X=i) = K(n,i) \cdot p^i \cdot (1-p)^{n-i}$$

fordi vi har $K(n,i)$ forskellige måder at arrangere vores succes'er og fiaskoer på, så der netop er i succes'er (vi skal udvælge i af de n pladser til vores succes'er). Hvert af disse udfald med i succeser har sandsynligheden $p^i \cdot (1-p)^{n-i}$ fordi vi, når vi siger at vi vil have netop i succes'er samtidig siger, at vi skal have netop $(n-i)$ fiaskoer.

Binomialkvotient

Bemærkning: Der argumenteres ikke for formlen i bogen. Nedenfor gives nogle konkrete eksempler på hvorfor formlen passer. Når man har forstået det, kan man faktisk godt forklare formlen generelt. Det er tilstrækkeligt, at i forklare det ud fra et taleksempel. Antallet af måder vi kan udvælge en r-del-mængde af en n-mængde beregnes ved følgende formel:

$$K(n,r) = \frac{n \cdot (n-1) \cdot (n-2) \cdot \dots}{r \cdot (r-1) \cdot (r-2) \cdot \dots \cdot 2 \cdot 1} = \frac{n!}{r! \cdot (n-r)!}$$

Tekst: Det står nedenfor.

1 - Ordnete delmængder og (ikke-ordnete) delmængder:

Vi har i det foregående talt om antallet af 3-del-mængder. Nu indfører vi begrebet antal ordnede 3-del-mængder. I de ordnede 3-del-mængder skelner vi mellem rækkefølgen.

Fordelen ved at betragte de ordnede delmængder er, at de er lettere at regne på.

Antallet af ordnede 3-del-mængder af en 7-mængde er ikke så svært at bestemme. Vi skal udfylde tre pladser med elementer fra mængden {a,b,c,d,e,f,g}:

--	--	--

På første plads er der 7 muligheder. Anden plads kan udfyldes på 6 måder når 1. plads er udfyldt og 3. plads på 5. måder.

Antallet at ordnede 3-del-mængder af en 7-mængde er

$$7 \cdot 6 \cdot 5 = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1} = \frac{7!}{(7-3)!}$$

Tilsvarende er antal ordnede 4-del-mængde af en 6-del-mængde

$$6 \cdot 5 \cdot 4 \cdot 3 = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = \frac{6!}{(6-4)!}$$

Vi kan altså udvælge en ordnet r-del-mængde af en n-mængde på $\frac{n!}{(n-r)!} = n \cdot (n-1) \cdot \dots$ måder

2 - Udregning af K(4,3)

Betragter vi mængden {a,b,c,d} så har vi i alt

$$4 \cdot 3 \cdot 2 = 24 \text{ eller } \frac{4!}{(4-3)!} = 24$$

forskellige ordnede 3-del-mængder af {a,b,c,d}. Vi kan skrive disse 24 ordnede delmængder op:

abc acb bac bca cab cba abd adb bad bda dab dba acd adc cad cda dac dca bcd bdc cbd cdb dbc dcb

Men ser vi ovenfor, så er de første 6 ordnede delmængder forskellige rækkefølger af den samme delmængde. Hver delmængde kan ordnes på 3·2·1 måder. Hvis hver delmængde forekommer i 3! forskellige ordninger, så kan vi beregne antal delmængder K(4,3) ved

$$K(4,3) = \frac{\frac{4!}{(4-3)!}}{3!} = \frac{4!}{3! \cdot (4-3)!} = \frac{4 \cdot 3 \cdot 2}{3 \cdot 2 \cdot 1} = \frac{24}{6} = 4$$

Der er i alt 4 forskellige 3-del-mængder af {a,b,c,d} :abc abd acd bcd

3 – Udregning af K(7,4)

Hvis vi vil bestemme K(7,4) så udregner vi først antal ordnede 4-del-mængden af en 7-mængde. De kan vælges på

$$7 \cdot 6 \cdot 5 \cdot 4 = \frac{7!}{(7-4)!} \text{ måder}$$

Hver 4-del-mængde kan ordnes på 4! eller 4·3·2·1 forskellige måder.

Antal 4-del-mængder af 7-mængde bliver dermed $K(7,4) = \frac{7!}{(7-4)! \cdot 4!}$

4 - Formelt argument for binomialkoefficienter:

Vi kan udtage en ordnet r-del-mængde af en n-mængde på

$$\frac{n!}{(n-r)!} = n \cdot (n-1) \cdot (n-2) \cdot \dots \text{ måder}$$

Hver r-del-mængde kan ordnes på r! forskellige måder. Derfor bliver antallet af forskellige r-del-mængder (ikke-ordnede) af en n-mængde

$$K(n,r) = \frac{n \cdot (n-1) \cdot (n-2) \cdot \dots}{r \cdot (r-1) \cdot (r-2) \cdot \dots \cdot 2 \cdot 1} = \frac{n!}{r! \cdot (n-r)!}$$

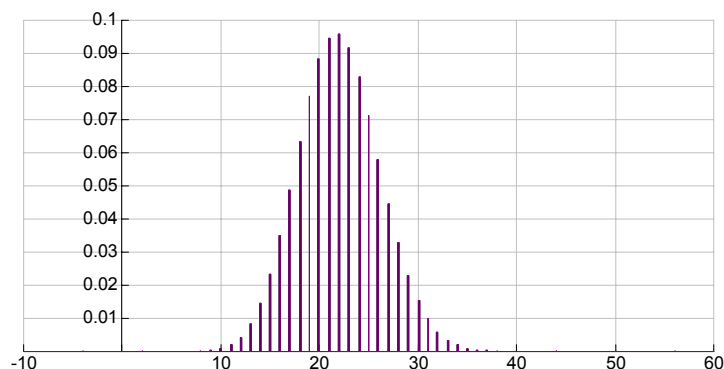
Binomialfordelingen nærmer sig normalfordelingen når n er stor

Tekst: II-arbejdsbog 134-137, II 174-181)

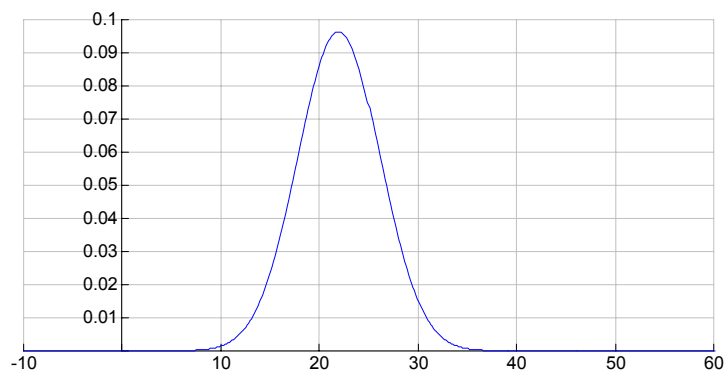
Binomialfordelingen opfører sig – når n bliver stor nok – som en Normalfordeling
Når n bliver stor i en $Bi(n, p)$ fordeling, så opfører den sig som en $N(n \cdot p, \sqrt{n \cdot p \cdot (1 - p)})$ -fordeling.

Du skal vide at normalfordelingen er en klokkeformet kontinuert fordeling, men ellers behøver du ikke vide ret meget mere om den.

$Bi(100, 0.22)$



$N(22, 4.14)$



Binomialfordelingstest og kritisk mængde

Bemærkning: Dette område er meget specielt og kun aktuelt hvis du har arbejdet med det tidligere.

Tekst: Du kan benytte nedenstående tekst. Du kan supplere med lærebogen II-162+165-174

1 – Kritisk mængde

Hvis vi har en hypotese om at andelen af socialdemokrater 22% og vi spørger 100 mennesker så vil vi forvente at omkring 22 af de spurgte vil være socialdemokrater.

Hvis der er mange flere eller meget færre der siger de er socialdemokrater forkaster vi vores hypotese og konkluderer at andelen er ændret. De tal der giver anledning til at forkaste hypotesen om at andelen uændret på 22% kaldes *den kritiske mængde*.

Oftentimes vælger man at sige at hvis en observation er så stor, at den ligger i de 2½% største eller hvis en observation er så lille, at den ligger i de 2½% mindste, så vil vi forkaste hypotesen.

2 – Kritisk mængde beregnet med TI

Hvis vores hypotese er at antal personer ud af 100 udspurgte, der svarer de er socialdemokrater er $bi(100, 0.22)$ -fordelt, så viser det sig at vores kritiske mængde for denne hypotese bliver $\{0, 1, 2, \dots, 12, 13\}$ U $\{30, 31, 32, \dots\}$

Vi kan med TI udregne fx

$$P(X \leq 14) = \text{binomcdf}(100, 0.22, 14) \quad .030452911$$

Her ser vi at sandsynligheden for at få 14 eller et antal der er endnu mindre er ca 3%. Vi befinder os altså ikke i den kritiske mængde.

Ud fra

$$\text{binomcdf}(100, 0.22, 13) \quad .01602022$$

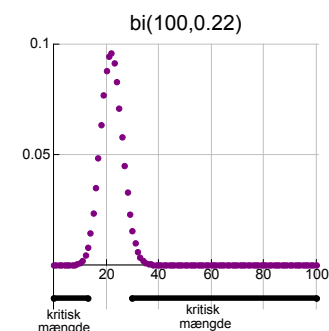
ses at 13 er grænsen for den kritiske mængde. Hvis vi får 13 eller et mindre tal er vi under de 2½% mindste observationer.

Det kan være fordi vi bare var ”uheldige”, men vi vælger at tro, at det snarere skyldes at opbakningen til socialdemokratiet er faldet.

I den anden ende ser vi at vi skal op på 30 eller flere for at nå op i de øverste 2½%

$$1 - \text{binomcdf}(100, 0.22, 30) \quad .023164253$$

$$1 - \text{binomcdf}(100, 0.22, 29) \quad .038554343$$



3 - Kritisk mængde bestemt ud fra middelværdi og spredning

Middelværdien er *den gennemsnitlige værdi* for en stokastisk variabel. Den kaldes normalt $E(X)$. Hvis $X \sim \text{bi}(n, p)$ så er middelværdien:

$$E(X) = n \cdot p$$

Spredningen er et slags mål for den *gennemsnitlige afvigelse fra middelværdien*. Den beregnes ved

$$\sigma(X) = \sqrt{n \cdot p \cdot (1 - p)}$$

og det er ikke lige til at forstå!

Man kan som tommelfinger regel få et godt bud på grænserne for den kritiske mængde ved at beregne $E(X) \pm 2 \cdot \sigma(X)$