Supplement til kapitel 7: Approksimationen til normalfordelingen, s. 136

Det er besværligt at regne med binomialfordelingen, og man vælger derfor ofte at bruge en approksimation med normalfordeling.

Man kan vise, at en binomialfordelt stokastisk variabel med parametre *n* og *p* kan tilnærmes med en normalfordeling med samme middelværdi og spredning, dvs. $\mu = n \cdot p$ og $\sigma = \sqrt{n \cdot p \cdot (1-p)}$.

Ved beregning af forkastelsesområderne (se kapitel 8) for binomialfordelingen kan man ofte bruge følgende approksimation til normalfordelingen.

Sætning 1 - approksimation med normalfordelingen

Hvis X er en binomialfordelt stokastisk variabel med parametre p og n, da vil følgende to sandsynligheder være omtrent lige store for vilkårlige a og b mellem 0 og n:

$$P\left(a \leq X \leq b\right)$$

Og:

$$P\left(\frac{a-\frac{1}{2}-n \cdot p}{\sqrt{n \cdot p \cdot \left(1-p\right)}} \leq Z \leq \frac{b+\frac{1}{2}-n \cdot p}{\sqrt{n \cdot p \cdot \left(1-p\right)}}\right)$$

Hvor Z er standardiseret normalfordelt *N(0,1)*, se side 154-155 i Statistik.

Approksimationen er gyldig for store værdier af n. Man kan nogenlunde stole på approksimationen, hvis følgende to betingelser er opfyldt:

$$(1) \qquad n > 30$$

og:

$$(2) \qquad n \cdot p \cdot \left(1-p\right) > 9$$

Betingelse (1) sikrer, at vi undgår situationer, hvor binomialfordelingen er meget skæv og derfor ikke kan approksimeres ved en normalfordeling.

Sætningen ovenfor kan også formuleres sådan, at binomialfordelingen og normalfordelingen ligner hinanden for store værdier af n.

For at se dette omskriver vi udtrykket i sætningen til:

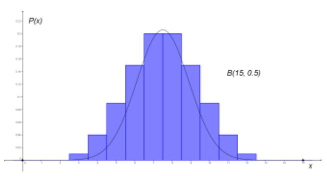
$$P\left(a-\tfrac{1}{2}<Z\cdot\sqrt{n\cdot p\cdot(1-p)}+n\cdot p<b+\tfrac{1}{2}\right)$$

I formlen ovenfor genkender vi de to udtryk $\mu=n\cdot p$ og $\sigma=\sqrt{n\cdot p\cdot(1-p)}$ for middelværdi og spredning i binomialfordelingen. Det er disse værdier, vi tilnærmer til normalfordelingens parametre μ og σ. Dette resultat var allerede opdaget af de Moivre i 1733 (se kapitel 8, afsnit 2). Den halve, som indgår i formlen, er der for at gøre approksimationen bedre. Man lægger 0,5 til/trækker 0,5 i siderne, idet binomialfordelingen er en diskret fordeling, mens normalfordelingen er en kontinuert fordeling. Det kaldes kontinuitetskorrektionen.

Fra normalfordelingen (kapitel 7, side 154-155) ved vi, at hvis X er normalfordelt med parametrene (μ,σ), da er $Z=\dfrac{X-\mu}{\sigma}$ normalfordelt med parametrene $(0,1)$. Heraf ses, at hvis Z er standardnormalfordelt, da er $U=Z\cdot\sqrt{n\cdot p\cdot(1-p)}+n\cdot p$ normalfordelt med parametre $n\cdot p$ og $\sqrt{n\cdot p\cdot(1-p)}$.

Sætningen udtrykker altså, at binomialfordelingen ligner normalfordelingen med parametre $n\cdot p$ og $\sqrt{n\cdot p\cdot(1-p)}$, når betingelserne (1) og (2) er opfyldt.

Man kan illustrere at de to fordelinger ligner hinanden ved at tegne deres frekvensfordelinger. F.eks. for n = 15 og p = 0.5.



Sammenligning af binomialfordelingen og normalfordelingen for $n=15$ og $p=0,5$.

Øvelse 1
Beregn hvilke værdier der er anvendt for normalfordelingens parametre.

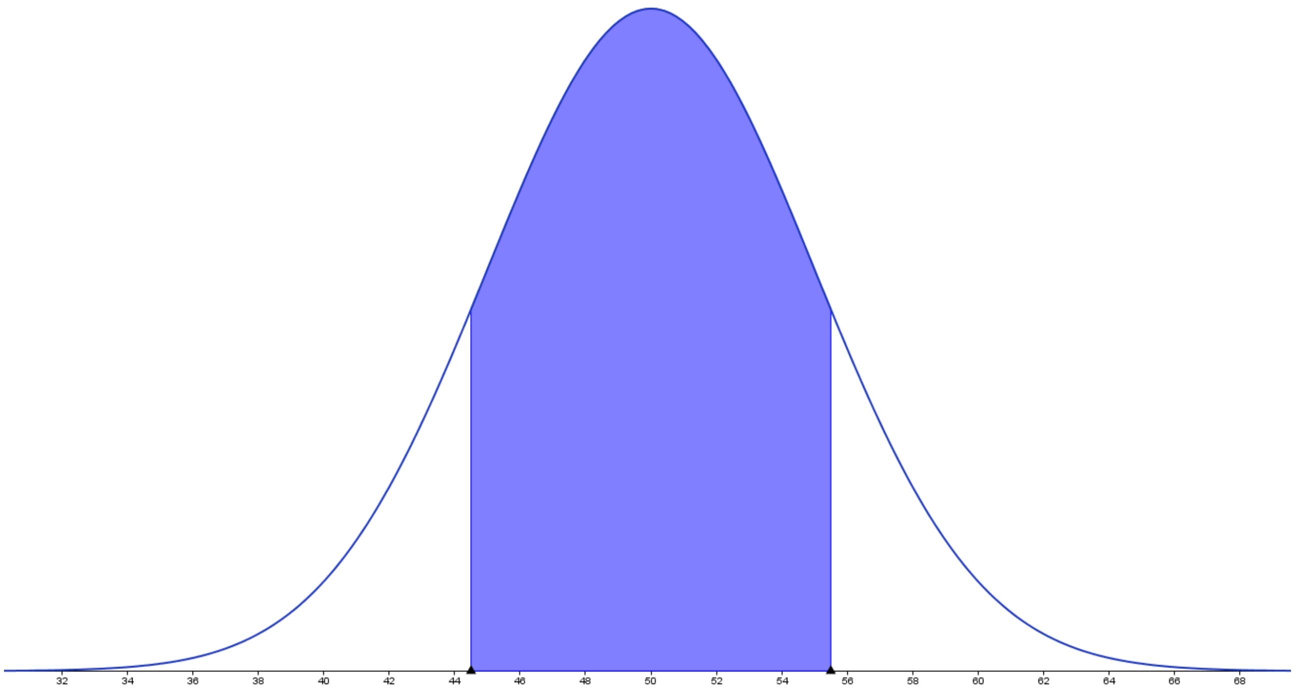
Vi illustrerer anvendelsen af sætning 1 med nogle eksempler.

Eksempel 1: Møntkast

Da det er besværligt at udregne binomialfordelingen, når antalsparameteren n er stor, vil vi bruge approksimationen til normalfordelingen. Vi tænker på eksperimentet, hvor vi kaster en mønt 100 gange, og den stokastiske variabel U tæller antallet af krone. Så hvis vi sætter $n=100$ og $p=\frac{1}{2}$, finder vi:

$$\mu = n\cdot p = 100\cdot\tfrac{1}{2} = 50 \text{ og } \sigma = \sqrt{n\cdot p\cdot(1-p)} = \sqrt{100\cdot\tfrac{1}{2}\cdot\tfrac{1}{2}} = \sqrt{25} = 5.$$

Vi ønsker at finde sandsynligheden $P(45 \le \text{antal krone} \le 55)$. Men vi vil ikke beregne værdien ved at benytte binomialfordelingen, men bruger approksimationen.



Vi skal finde det interval, der svarer til mindst 45 antal krone og højst 55 antal krone, og inkluderer derfor intervallet 44,5-45,5, som ligger lige langt fra 45, og 54,5- 55,5, som også ligger lige langt fra 55, altså en halv til hver side af de to tal 45 og 55. Derfor bliver opgaven ved hjælp af normalfordelingen at finde $P(44{,}5 \le \text{antal krone} \le 55{,}5)$.

Til sidst omskriver vi til den kendte transformation (se kapitel 7):

$$P\left(\frac{45-\frac{1}{2}-50}{5} \le Z \le \frac{55+\frac{1}{2}-50}{5}\right) = P(-1{,}1 \le Z \le 1{,}1) = \Phi(1{,}1) - \Phi(-1{,}1) = 0{,}7287$$

Dvs. sandsynligheden for at få mellem 45 og 55 gange krone er 72,87 %.

Den eksakte værdi kan beregnes med Maple:

$with(Statistics):$

$X := RandomVariable(Binomial(100, 0.5))$

$$\_R$$

$CDF(X, 55) - CDF(X, 44)$

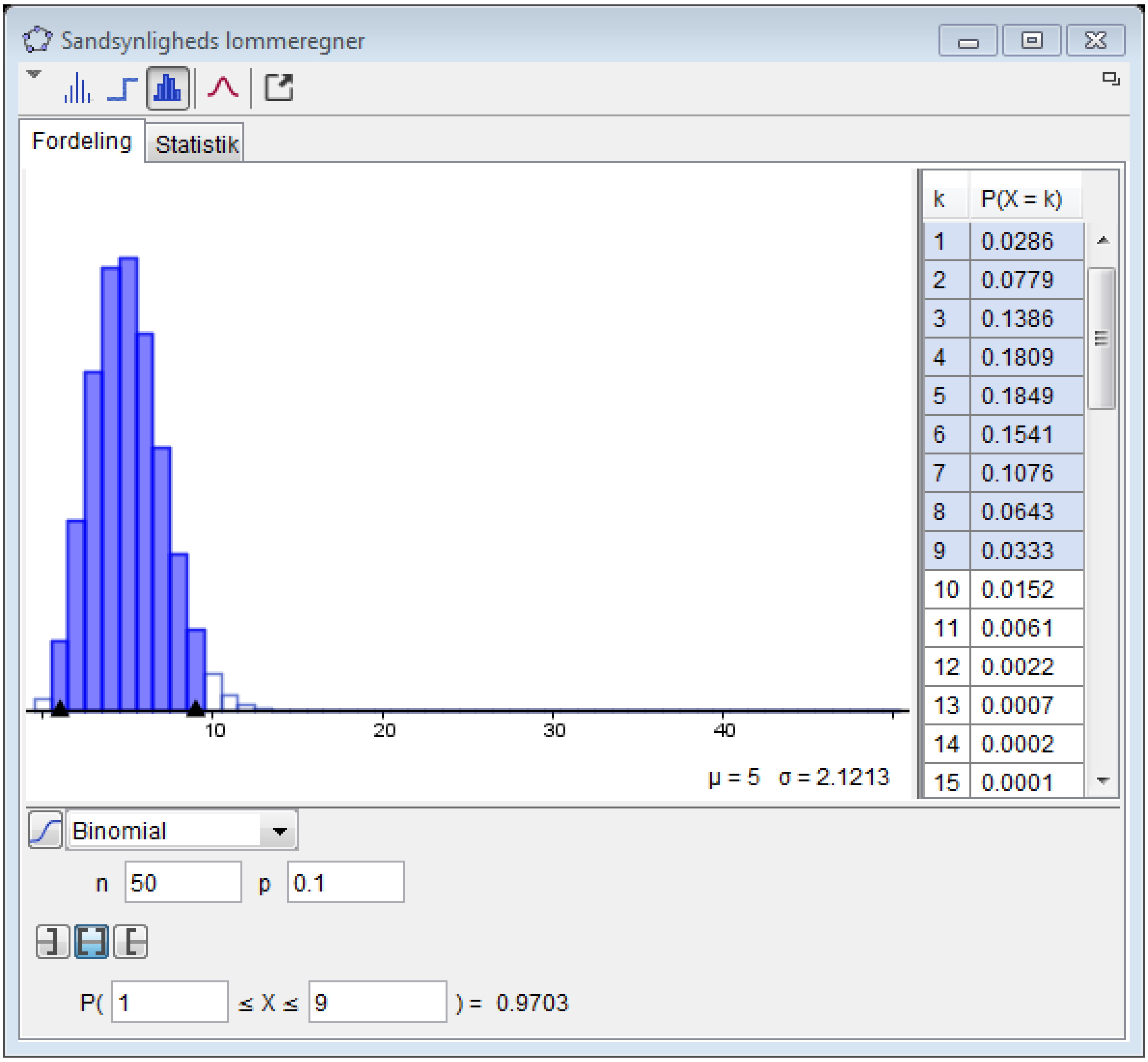
$$0.7287469760$$

Med fire betydende cifre fås samme resultat.

Eksempel 2

Ved en opinionsundersøgelse per telefon skal vi udvælge 10 % af en befolkning tilfældigt. Vi vedtager at vælge alle, hvis telefonnummer ender med tallet 3. Man vælger tilfældigt 50 numre i telefonbogen og finder 9, der ender på et tretal. Vi skal her teste hypotesen $p = 0{,}1$ og $n = 50$.

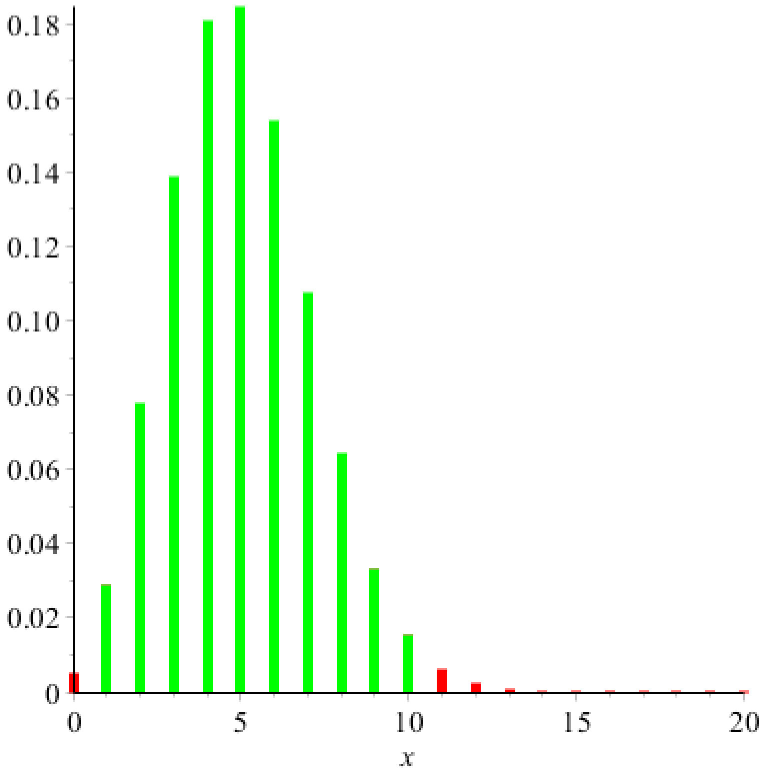
Ved brug af et CAS-værktøj finder vi, at x = 1 og x = 10 med en vis tilnærmelse afgrænser de to 2,5 % -haler i en binomialfordeling med $p = 0{,}1$ og $n = 50$. Da x = 9 falder inden for disse to grænser, kan vi ikke afvise, at proceduren faktisk giver os ca. 10 % af populationen.



Plot i GeoGebra af binomialfordelingen med $p = 0{,}1$ og $n = 50$. Det blå område (x = 1 til 9) svarer til 97 %.

Øvelse 1
Check selv efter med sandsynlighedslommeregneren i GeoGebra at intervallet x = 1 til 9 er det rigtige interval, idet det forkastelsesområdet (det ikke-blå) skal være så tæt på 5 % som muligt, men ikke større end 5 %.

I Maple ser plottet således ud:



Det grønne er acceptområdet, det røde er forkastelsesområdet.

Vi vil nu beregne grænserne ved brug af sætning 1 ovenfor. Opgaven er altså at finde forkastelsesområdet for hypotesen $p=0{,}1$, når $n=50$. Eller anderledes sagt at bestemme to tal a_1 og b_1, så vi har:

$$P(X \le a_1)=0{,}025 \text{ og } P(X \ge b_1)=0{,}025$$

Der gælder $P(X \le a_1)=\Phi\left(\frac{a_1+0{,}5-n\cdot p}{\sqrt{n\cdot p\cdot(1-p)}}\right)=0{,}025$, hvor Φ er standardnormalfordelingen (se kapitel 7). For Φ-fordelingen ved vi imidlertid, at -1,96 afgrænser de nederste 2,5 % af fordelingen.
Da $\Phi(-1{,}96)=0{,}025$, følger heraf at:

$$\frac{a_1+0{,}5-n\cdot p}{\sqrt{n\cdot p\cdot(1-p)}}=-1{,}96 \text{ eller } a_1=n\cdot p-\left(0{,}5+1{,}96\cdot\sqrt{n\cdot p\cdot(1-p)}\right)$$

På samme måde finder vi, at:

$$P(X \ge b_1)=1-\Phi\left(\frac{b_1-0{,}5-n\cdot p}{\sqrt{n\cdot p\cdot(1-p)}}\right)=0{,}025$$

Men da 1,96 afgrænser de øverste 2,5 % af Φ-fordelingen følger heraf, at:

$$b_1 = n \cdot p + \left(0,5 + 1,96 \cdot \sqrt{n \cdot p \cdot (1-p)}\right).$$

Ved at indsætte vores værdier, finder vi det kritiske område, som indeholder *x*-værdier mindre end eller lig med $a_1 = 5 - \left(0,5 + 1,96 \cdot \sqrt{4,5}\right) = 0,34$ eller større end eller lig med $b_1 = 5 + \left(0,5 + 1,96 \cdot \sqrt{4,5}\right) = 9.66$.

En sammenligning med den første del af eksemplet viser at approksimationen til normalfordeling er meget nøjagtig. Der fik vi $a = 1$ og $b = 10$.

Eksempel 3: Folketingsvalg

Ved et folketingsvalg stemmer 17,8 % af vælgerne på et bestemt parti. En opinionsundersøgelse baseret på 1.500 personer viser senere, at 20,6 % ville stemme på partiet, hvis der var valg den næste dag.
Er fremgangen på 2,8 % en reel fremgang?

En måde at besvare dette spørgsmål på, er at teste hypotesen $p = 0,178$ i en binomialfordeling med $n = 1500$. Antal personer, der påstår, at de vil stemme på partiet, kan nemlig opfattes som heldige forsøg i en serie på 1.500 forsøg, hvor chancen for et heldigt forsøg er lig med procentdelen af vælgere, der faktisk ville stemme på partiet ved et forstående valg.

Med $n = 1500$ og $p = 0,178$ bliver de to grænser lig med:

$$a_1 = 1500 \cdot 0,178 - \left(0,5 + 1,96 \cdot \sqrt{1500 \cdot 0,178 \cdot 0,822}\right) = 237,463$$

$$b_1 = 1500 \cdot 0,178 + \left(0,5 + 1,96 \cdot \sqrt{1500 \cdot 0,178 \cdot 0,822}\right) = 296,537$$

Vi må altså afvise hypotesen $p = 0,178$, hvis $x \geq 296,5$. Den observerede procentdel 20,6 % af 1.500 personer svarer til $x = 1500 \cdot 0,206 = 309$, og vi må forkaste hypotesen. Det observerede procenttal tyder således på, at partiet har haft en vis fremgang siden valget.

Opgaver

Opgave 1

Sammenlign de eksakte sandsynligheder for $P(X \geq a)$, når a = 13, 14, 15, 16 og 17, med normalfordelingens approksimation, når X er binomialfordelt med $n = 30$ og $p = 0,5$.