MAT0100V
Sannsynlighetsregning og kombinatorikk

Oppgaver om
- **Binomisk og hypergeometrisk fordeling**
- **Forventning, varians og standardavvik**
- **Tilnærming av binomiske sannsynligheter**
- **Konfidensintervall**

Ørnulf Borgan
Matematisk institutt
Universitetet i Oslo

Hypergeometrisk fordeling

Vi har følgende situasjon:

- Vi har en mengde med *N* elementer
- Elementene i mengden kan deles inn i to delmengder D og $\bar{D}$
  Det er *m* elementer i *D* og *N* – *m* elementer i $\bar{D}$
- Vi trekker tilfeldig *n* elementer fra mengden uten tilbakelegging

La *X* være antall elementer vi trekker fra *D*

$$P(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}$$

Binomisk fordeling

Vi har følgende situasjon:

- Vi gjør *n* forsøk
- I hvert forsøk er det to muligheter:
  Enten inntreffer en bestemt begivenhet *S*
  ellers så gjør den ikke det
- I hvert forsøk er sannsynligheten lik *p* for at *S* skal inntreffe
- Forsøkene er uavhengige

X er antall ganger *S* inntreffer i de *n* forsøkene

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Oppgave 85 I en urne er det tre røde kuler og to hvite kuler. To kuler trekkes tilfeldig ut, henholdsvis uten og med tilbakelegging. La X angi antall røde kuler vi trekker.

a) Hva gir binomisk fordeling for X, trekking med eller uten tilbakelegging?

b) Beregn $P(X = x)$ for $x = 0, 1, 2$ både for trekning med og uten tilbakelegging.

a) Trekning uten tilbakelegging: hypergeometrisk
Trekning med tilbakelegging: binomisk

b) Uten tilbakelegging: $P(X = k) = \frac{\binom{3}{k} \cdot \binom{2}{2-k}}{\binom{5}{2}}$

Med tilbakelegging: $P(X = k) = \binom{2}{k} \left(\frac{3}{5}\right)^k \left(\frac{2}{5}\right)^{2-k}$

	k	0	1	2
Uten	P(*X* = *k*)	0.10	0.60	0.30
Med	P(*X* = *k*)	0.16	0.48	0.36

Anta så at det er 300 røde kuler og 200 hvite kuler i urnen. Vi lar fortsatt X angi antall røde kuler i to trekninger.

c) Beregn også nå $P(X = x)$ for $x = 0, 1, 2$ for trekning med og uten tilbakelegging. Sammenlign med det du fant i punkt a.

c)

Uten tilbakelegging: $P(X = k) = \dfrac{\binom{300}{k} \cdot \binom{200}{2-k}}{\binom{500}{2}}$

Med tilbakelegging: $P(X = k) = \binom{2}{k} \left(\frac{300}{500}\right)^{k} \left(\frac{200}{500}\right)^{2-k}$

Uten (b)	P(X = k)	0.10	0.60	0.30
Uten (c)	P(X = k)	0.1595	0.4810	0.3595
Med	P(X = k)	0.16	0.48	0.36

Resultatet i punkt c illustrerer at hvis vi trekker et (relativt sett) lite utvalg fra en (relativt sett) stor populasjon, så er det ubetydelig forskjell på trekning med og uten tilbakelegging.

Ved en meningsmåling om holdningen til norsk EU-medlemskap trekkes det tilfeldig (og uten tilbakelegging) et utvalg på 1000 av befolkningen over 18 år. La X være antallet i utvalget som er mot norsk EU-medlemskap.

d) Hvorfor er det rimelig å anta at X er binomisk fordelt med $n = 1000$ og p lik (den ukjente) andelen av den voksne befolkningen som er mot norsk EU-medlemskap?

d) Utvalget vi trekker er lite i forhold til hele populasjonen. Derfor er det rimelig å anta at *X* er binomisk fordelt.

Oppgave 86 Et politisk parti har ved et bestemt tidspunkt støtte av 25% av befolkningen. Tjue personer blir trukket ut tilfeldig. La X angi hvor mange av de uttrukne som støtter partiet.

a) Forklar hvorfor vi kan regne som om X er binomisk fordelt med $n = 20$ og $p = 0.25$.

b) Skriv opp en formel for sannsynlighetsfordelingen til X.

c) Bruk formelen til å finne sannsynligheten for at (i) akkurat 3 støtter partiet; (ii) akkurat 4 støtter partiet; (iii) akkurat 5 støtter partiet.

a) Jf. forrige oppgave.

b) $P(X = k) = \binom{20}{k} \cdot 0.25^{k} \cdot 0.75^{20-k}$

c) $P(X = 3) = \binom{20}{3} \cdot 0.25^{3} \cdot 0.75^{20-3}$

$= 1140 \cdot 0.25^{3} \cdot 0.75^{17} = 0.134$

$P(X = 4) = 0.190$

$P(X = 5) = 0.202$

En tilfeldig variabel *X* har mulige verdier $x_1, x_2, \ldots, x_m$. Da er forventningsverdien

$$\mu = E(X) = x_1 \cdot P(X = x_1) + \ldots + x_m \cdot P(X = x_m)$$

En tilfeldig variabel *X* har mulige verdier $x_1, x_2, \ldots, x_m$ og forventningsverdi μ

Da er variansen

$$\sigma^2 = \mathrm{Var}(X)$$
$$= (x_1 - \mu)^2 \cdot P(X = x_1) + \ldots + (x_m - \mu)^2 \cdot P(X = x_m)$$

Standardavviket til en tilfeldig variabel *X* er gitt ved

$$\sigma = SD(X) = \sqrt{\mathrm{Var}(X)}$$

Oppgave 90 En rekke personer forsikrer sine sykler i et forsikringsselskap. Vi antar for enkelhets skyld at en person bare kan få erstatning av selskapet av to grunner: *(i)* sykkelen blir stjålet og dukker ikke opp igjen, eller *(ii)* den blir stjålet, men kommer senere til rette delvis ramponert. I det første tilfellet får den forsikrede en erstatning på 3500 kroner (etter at egenandel er trukket fra), mens han i det andre tilfellet får 1000 kroner. Vi antar at sannsynligheten for *(i)* er 2%, mens sannsynligheten for *(ii)* er 5%.

Oppgave 101 Se på oppgave 90. Finn variansen og standardavviket til den erstatningen en forsikret sykkeleier vil få fra selskapet i løpet av ett år. Hvilken benevning har variansen og standardavviket?

X er erstatning for en tilfeldig valgt sykkeleier

$$E(X) = 0 \cdot 0.93 + 1000 \cdot 0.05 + 3500 \cdot 0.02 = 120 \text{ kroner}$$

$$\text{Var}(X)$$

$$= (0 - 120)^2 \cdot 0.93 + (1000 - 120)^2 \cdot 0.05 + (3500 - 120)^2 \cdot 0.02$$

$$= 280600 \text{ kroner}^2$$

$$SD(X) = \sqrt{280600} = 529.72 \text{ kroner}$$

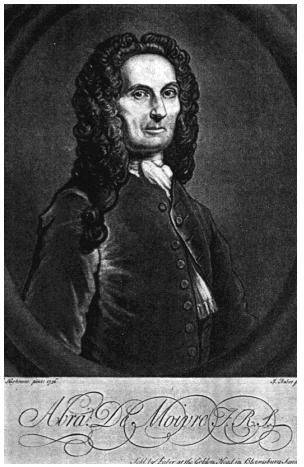
Tilnærming av binomiske sannsynligheter

Tidligere var det vanskelig å bruke formelen for binomisk fordeling til å regne ut sannsynligheter når n er stor

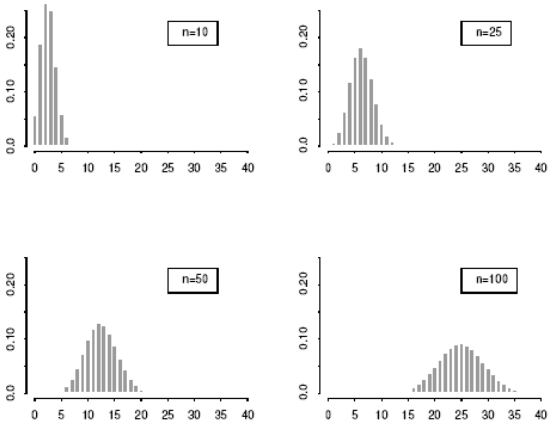
Alt i 1733 viste Abraham de Moivre hvordan en kan finne tilnærmingsverdier for binomiske sannsynligheter

Selv om det nå er enklere å bestemme binomiske sannsynligheter, er denne tilnærmelsen fortsatt viktig

(www.york.ac.uk/depts/maths/histstat)



Binomisk fordeling for p = 0.25 og n = 10, 25, 50, 100



Fordelingen forskyves mot høyre og blir mer «spredt ut» når *n* øker

For å finne en tilnærming «forskyver» vi fordelingene slik at de får « tyngdepunktet» i origo, og vi «skalerer» dem slik at de får samme spredning

Vi ser derfor på den standardiserte variabelen

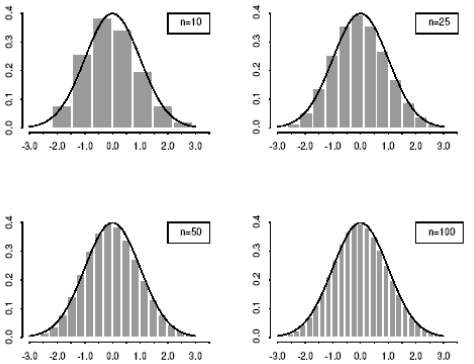
$$Z = \frac{X - E(X)}{SD(X)} = \frac{X - np}{\sqrt{np(1-p)}}$$

Vi har at E(Z) = 0 og SD(Z) = 1

Vi merker oss at hvis $X = k$ så er $Z = \frac{k - np}{\sqrt{np(1-p)}}$

Vi får derfor fordelingen til Z av fordelingen til X

Stolpediagram for fordelingen til Z



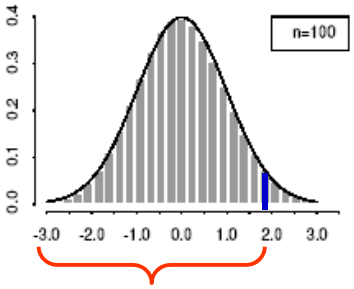
Arealet av en stolpe svarer til sannsynligheten for at Z får den aktuelle verdien

Stolpediagrammene nærmer seg standard-normalfordelingsfunksjonen $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Vil bruke de Moivres tilnærming til å finne $P(X \le 33)$ når n = 100 og p = 0.25

Vi merker oss at

$$P(X \le 33) = P\left(\frac{X - 100 \cdot 0.25}{\sqrt{100 \cdot 0.25 \cdot 0.75}} \le \frac{33 - 100 \cdot 0.25}{\sqrt{100 \cdot 0.25 \cdot 0.75}}\right) = P(Z \le 1.85)$$



Vi skal egentlig summere arealene av alle søylene til venstre for 1.85

Summen av arealene av søylene er omtrent like stor som arealet under f(x) til venstre for 1.85

Arealet under standardnormalfordelings-funksjonen til venstre for **1.85** finner vi av tabellen bak i kompendiet:

z	Siste desimal i z									
	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

De Moivres tilnærming gir at

$$P(X \le 33) = P(Z \le 1.85) \approx 0.968$$

Oppgave 111 Ved Stortingsvalget i 2013 stemte 30.8% av velgerne på Arbeiderpartiet. Tenk deg at du var med på å gjennomføre en valgdagsmåling det året, der et tilfeldig utvalg på 5000 velgere ble spurt om hvilket parti de nettopp stemte på. Hva er sannsynligheten for at a) høyst 1500 av dem hadde stemt Arbeiderpartiet; b) minst 1600 av dem hadde stemt Arbeiderpartiet?

La X være antall som ville ha stemt på Ap

X er binomisk fordelt med n = 5000 og p = 0.308

a) Ved de Moivres tilnærming finner vi at

$$P(X \le 1500) = P\left(\frac{X - 5000 \cdot 0.308}{\sqrt{5000 \cdot 0.308 \cdot 0.692}} \le \frac{1500 - 5000 \cdot 0.308}{\sqrt{5000 \cdot 0.308 \cdot 0.692}}\right)$$

$$= P(Z \le -1.23) \approx 0.109$$

b) $$P(X \ge 1600) = P\left(\frac{X - 5000 \cdot 0.308}{\sqrt{5000 \cdot 0.308 \cdot 0.692}} \ge \frac{1600 - 5000 \cdot 0.308}{\sqrt{5000 \cdot 0.308 \cdot 0.692}}\right)$$

$$= P(Z \ge 1.84) \approx 1 - 0.967 = 0.033$$

Estimering og konfidensintervall

Vi er interessert i å anslå («estimere») verdien av *p* ut fra resultatet av et forsøk, og også å si noe om hvor presist anslaget er

Vi ser på en stor populasjon der en ukjent andel *p* har et bestemt «kjennetegn»

Vi trekker et tilfeldig utvalg på *n* individer fra populasjonen. Størrelsen av utvalget er liten i forhold til størrelsen av hele populasjonen

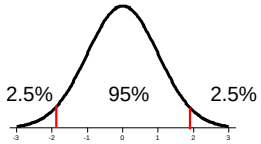
La X være antall i utvalget som har kjennetegnet

Vi kan regne som om X er binomisk fordelt med *p* lik den ukjente andelen i populasjonen som har kjennetegnet

Til å anslå («estimere») *p* bruker vi andelen i utvalget som har kjennetegnet, dvs. $\hat{p} = X / n$

Av de Moivres resultat finner vi at

$$P\left(-1.96 \le \frac{X-np}{\sqrt{np(1-p)}} \le 1.96\right) \approx 0.95$$



Nå er $\frac{X-np}{\sqrt{np(1-p)}} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

Det følger at

$$P\left(-1.96 \le \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \le 1.96\right) \approx 0.95$$

Kan vise at vi kan erstatte *p* med $\hat{p}$ i nevneren:

$$P\left(-1.96 \le \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le 1.96\right) \approx 0.95$$

Ulikhetene kan omformes slik at vi får *p* alene i midten:

$$P\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 0.95$$

Det er altså tilnærmet 95% sannsynlig at undersøkelsen vil gi et resultat som er slik at *p* blir liggende i konfidensintervallet:

$$\left[\ \hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ \right]$$

Oppgave 114 En frøprodusent ønsker å bestemme spireprosenten av en bestemt type Kornblomstfrø. For dette formålet vil han så 500 frø og registrere hvor mange av dem som spirer.

a) Gjør deg noen tanker om hvordan frøprodusenten bør velge ut de 500 frøene.

b) Anta at 337 av frøene spirer. Gi estimatet for spireprosenten og bestem et 95% konfidensintervall for denne.

a) Hvis det produseres frø på flere jorder, bør produsenten velge frø fra tilfeldige steder på alle jordene og blande dem godt før de 500 frøene velges ut

b) Vi har $n = 500$ og $\hat{p} = 337/500 = 0.674$

Et 95% konfidensintervall for spireevnen er

$$\left[\ 0.674 - 1.96\sqrt{\frac{0.674\cdot(1-0.674)}{500}}\ ,\ 0.674 + 1.96\sqrt{\frac{0.674\cdot(1-0.674)}{500}}\ \right]$$

dvs. $[\ 0.633\ ,\ 0.715\ ]$

Oppgave 116 I 2010 var det i Norge 60 608 fødsler hvorav 1002 resulterte i flerfødsler (tvillinger, trillinger, osv.). De tilsvarende tallene for 1987 var (ca.) 53 500 og 599. Bestem (tilnærmete) 95% konfidensintervall for sannsynligheten for at et svangerskap skal resultere i en flerfødsel for hver av disse to årene. Er det noen grunn til å tro at det har vært en reell endring i sannsynligheten for at et svangerskap skal resultere i en flerfødsel fra 1987 til 2010?

2010: Vi har $n = 60608$ og $\hat{p} = 1002 / 60608 = 0.0165$

Et 95% konfidensintervall er

$$\left[\ 0.0165 - 1.96\sqrt{\frac{0.0165 \cdot 0.9835}{60608}}\ ,\ \ 0.0165 + 1.96\sqrt{\frac{0.0165 \cdot 0.9835}{60608}}\ \right]$$

dvs. $[\ 0.0155\ ,\ 0.0175\ ]$

1987: Vi har $n = 53500$ og $\hat{p} = 599 / 53500 = 0.0111$

$$\left[\ 0.0111 - 1.96\sqrt{\frac{0.0111 \cdot 0.9889}{53500}}\ ,\ \ 0.0111 + 1.96\sqrt{\frac{0.0111 \cdot 0.9889}{53500}}\ \right]$$

dvs. $[\ 0.0103\ ,\ 0.0121\ ]$

Oppgave 119 På British Museum i London fins det mange gamle terninger, blant annet en marmorterning fra romertiden. Tidsskriftet Biometrika ga i 1955 resultatet av 204 kast med denne marmorterningen. Av de 204 kastene ga 54 av kastene en sekser. Er det grunn til å tro at sannsynligheten for å få sekser med marmorterningen avviker fra 1/6?

Vi lager et 95% konfidensintervall for sannsynligheten for sekser

Vi har $n = 204$ og $\hat{p} = 54 / 204 = 0.265$

Et 95% konfidensintervall er

$$\left[\ 0.265 - 1.96\sqrt{\frac{0.265 \cdot 0.735}{204}}\ ,\ \ 0.0165 + 1.96\sqrt{\frac{0.265 \cdot 0.735}{204}}\ \right]$$

dvs. $[\ 0.204\ ,\ 0.325\ ]$