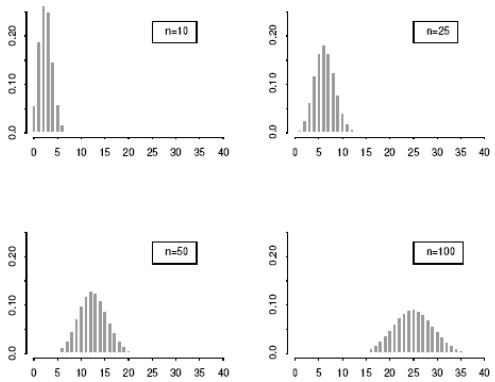
MAT0100V
Sannsynlighetsregning og kombinatorikk

Tilnærming av binomiske sannsynligheter
Konfidensintervall

Ørnulf Borgan
Matematisk institutt
Universitetet i Oslo

Tilnærming av binomiske sannsynligheter

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

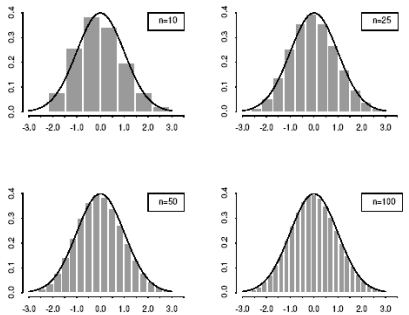
$$Z = \frac{X - E(X)}{SD(X)} = \frac{X - np}{\sqrt{np(1-p)}}$$

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

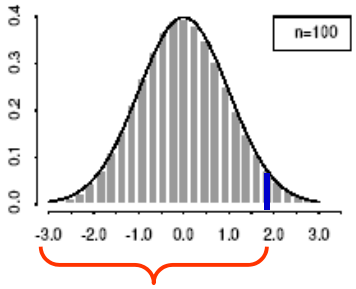
Stolpediagrammene nærmer seg standard-normalfordelingsfunksjonen $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Dette er de Moivres tilnærming for binomisk fordeling

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

Vi merker oss at

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



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

Vi skal egentlig summere arealene av alle søylene 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

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

Eksempel 10.3: Vi tenker oss at Arbeiderpartiet på et tidspunkt har oppslutning av 32.0% av velgerne

Et meningsmålingsinstitutt spør et tilfeldig utvalg på 1000 personer over 18 år hvilket parti de ville stemt på hvis det hadde vært valg

Hva er sannsynligheten for at mellom 300 og 340 av dem ville ha stemt på Arbeiderpartiet?

Med andre ord: hva er sannsynligheten for at Arbeiderpartiets oppslutning på menings-målingen vil bli mellom 30.0% og 34.0% ?

La X være antallet av de spurte som ville ha stemt på Arbeiderpartiet

Siden det trekkes uten tilbakelegging, er strengt tatt X hypergeometrisk fordelt

Men da antallet som trekkes ut er lite i forhold til antall over 18 år i hele befolkningen, kan vi regne som om X er binomisk fordelt med n = 1000 og p = 0.32

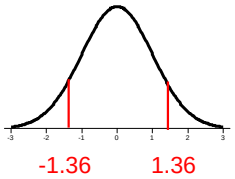
Nå har vi at

$$P(300 \le X \le 340)$$

$$= P\left(\frac{300-1000\cdot 0.32}{\sqrt{1000\cdot 0.32\cdot 0.68}} \le Z \le \frac{340-1000\cdot 0.32}{\sqrt{1000\cdot 0.32\cdot 0.68}}\right)$$

$$= P(-1.36 \le Z \le 1.36)$$

Arealet under standard-normalfordeligsfunksjonen mellom -1.36 og 1.36 er lik arealet til venstre for 1.36 minus arealet til venstre for -1.36



	Siste desimal i z									
z	0	1	2	3	4	5	6	7	8	9
−3	0.0013	0.0010	0.0007	0.0005	0.0003	0.0002	0.0002	0.0001	0.0000	0.0000
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

De Moivres tilnærming gir at

$$P(300 \le X \le 340) = P(-1.36 \le Z \le 1.36)$$

$$\approx 0.913 - 0.087 = 0.826$$

Oppgave 112 Tenk deg at Høyre på et tidspunkt har oppslutning av 25.0% av velgerne. Et meningsmålingsinstitutt spør et tilfeldig utvalg på 1000 personer over 18 år hvilket parti de ville ha stemt på hvis det hadde vært stortingsvalg i morgen. Hva er sannsynligheten for at Høyre vil få en oppslutning på meningsmålingen som er mellom 23.0% og 27.0% ? (Vi forutsetter at alle de spurte ville ha stemt hvis det hadde vært valg.)

Løsning:

La X være antall som ville ha stemt på Høyre
X er binomisk fordelt med $n = 1000$ og $p = 0.25$.

Høyre får en oppslutning på meningsmålingen mellom 23.0% og 27.0% hvis X får en verdi mellom 230 og 270

Ved de Moivres tilnærming finner vi at

$$P(230 \le X \le 270) = P\left(\frac{230-1000\cdot 0.25}{\sqrt{1000\cdot 0.25\cdot 0.75}} \le Z \le \frac{270-1000\cdot 0.25}{\sqrt{1000\cdot 0.25\cdot 0.75}}\right)$$

$$= P(-1.46 \le Z \le 1.46) \approx 0.928 - 0.072 = 0.856$$

Utfordring (gitt på samlingen):

Hvor mange ganger må vi kaste et kronestykke for at sannsynligheten skal være 95 % for at den relative frekvensen for mynt skal være mellom 49.5% og 50.5% ?

Løsning:

La X være mynt vi får når vi kaster et kronestykke n ganger
X er binomisk fordelt med n forsøk og $p = 0.50$

Den relative frekvensen for mynt er mellom 49.5% og 50.5% hvis X får en verdi mellom $0.495n$ og $0.505n$

Ved de Moivres tilnærming finner vi at

$$P(0.495n \le X \le 0.505n)$$

$$= P\left(\frac{0.495n-0.500n}{\sqrt{n\cdot 0.50\cdot 0.50}} \le Z \le \frac{0.505n-0.500n}{\sqrt{n\cdot 0.50\cdot 0.50}}\right)$$

$$= P\left(\frac{-0.005n}{0.50\sqrt{n}} \le Z \le \frac{0.005n}{0.50\sqrt{n}}\right)$$

$$= P\left(-0.01\sqrt{n} \le Z \le 0.01\sqrt{n}\right)$$

Vi vil bestemme *n* slik at denne sannsynligheten blir (ca) 95%

Av tabellen bak i kompendiet finner vi at

$$P(-1.96 \le Z \le 1.96) \approx 0.95$$

Vi bestemmer dermed *n* av ligningen

$$0.01\sqrt{n} = 1.96$$

Det gir

$$n = \left(\frac{1.96}{0.01}\right)^2 = 38416$$

Sannsynlighetsregning og statistikk

Vi har sett på tilfeldige variabler og deres sannsynlighetsfordelinger. Det er en del av sannsynlighetsregningen

Vi vil nå se på hvordan sannsynlighetsregningen danner grunnlaget for statistiske metoder

Vi nøyer oss med å se på binomiske situasjoner

I sannsynlighetsregningen kjenner vi verdien til *p*

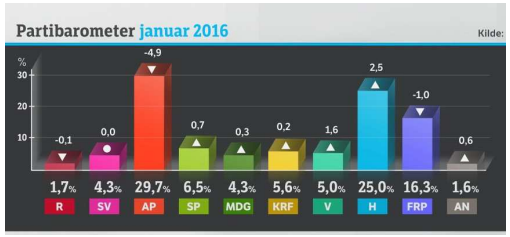
I statistikken gjør vi ikke det. Der er poenget nettopp å kunne si noe om verdien til *p* når vi har observert *X*

Estimering og konfidensintervall

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

Eksempel 11.1:

Av 721 personer som ville ha stemt hvis det hadde vært valg, ville 180 ha stemt på Høyre



Høyres oppslutning er 180/721=0.250, dvs. 25.0%

Hvor sikkert er dette anslaget?

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

I eksemplet er populasjonen alle over 18 år som ville ha stemt hvis det var valg, og kjennetegnet er at en person ville stemt på Høyre

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.

$$\hat{p} = \frac{X}{n}$$

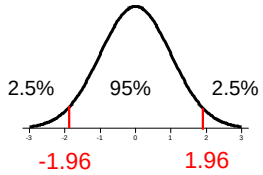
Merk at $\hat{p}$ («p hatt») er en tilfeldig variabel

I eksemplet fikk $\hat{p}$ verdien 0.250

For å kunne si noe om hvor presist et anslag er, må vi ta hensyn til hvor mye verdien av $\hat{p}$ vil variere fra undersøkelse til undersøkelse bare på grunn av tilfeldige variasjoner

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

Dermed

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

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:

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

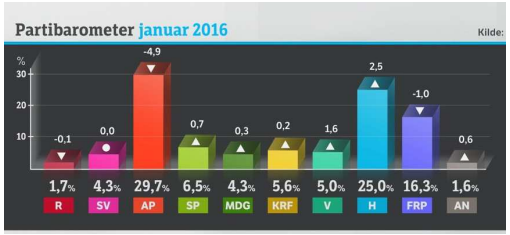
Det er altså tilnærmet 95% sannsynlig at undersøkelsen vil gi et resultat som er slik at *p* blir liggende i intervallet

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

Dette intervallet kaller vi et (tilnærmet) 95% konfidensintervall for *p*

Eksempel 11.2:

Vi ser igjen på meningsmålingen



Vårt estimat for Høyres oppslutning er:

$$\hat{p} = \frac{180}{721} = 0.250$$

95% konfidensintervall:

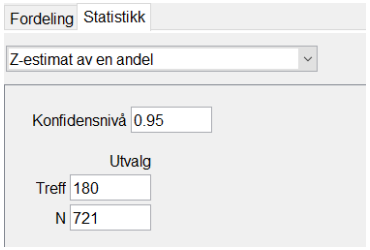
$$\left[\ 0.250 - 1.96\sqrt{\frac{0.250\cdot(1-0.250)}{721}}\ ,\ \ 0.250 + 1.96\sqrt{\frac{0.250\cdot(1-0.250)}{721}}\ \right]$$

Dvs.: $[\ 0.218\ ,\ 0.282\ ]$ (dette gir en «feilmargin»)

Vi kan bruke GeoGebra til å bestemme konfidensintervallet i eksempel 11.2.

Vi åpner da sannsynlighetskalkulatoren og velger «Statistikk» og «Z-estimat av en andel».

Så fyller vi inn slik det er vist til venstre nedenfor.
Da får vi konfidensintervallet slik det er vist til høyre nedenfor.



Z-estimat av en andel

Treff	180
N	721
SF	0.0161
Nedre grense	0.2181
Øvre grense	0.2812
Intervall	0.2497 ± 0.0316

Oppgave 117 Ved den politiske meningsmålingen for NRK for januar 2016 fikk Arbeiderpartiet en oppslutning på 29.7%, Fremskrittspartiet en oppslutning på 16.3%, Miljøpartiet de grønne en oppslutning på 4.3% og Rødt en oppslutning på 1.7%. Undersøkelsen er basert på intervju med 721 personer over 18 år som ville ha stemt hvis det hadde vært stortingsvalg.

a) Bestem et (tilnærmet) 95% konfidensintervall for oppslutningen om Arbeiderpartiet. Klargjør hvilke forutsetninger dette bygger på. Er disse forutsetningene noenlunde realistiske?

Løsning:

Vi har $n = 721$ og $\hat{p} = 0.297$

Et 95% konfidensintervall for Arbeiderpartiets oppslutning er

$$\left[\ 0.297 - 1.96\sqrt{\frac{0.297\cdot(1-0.291)}{721}}\ ,\ \ 0.297 + 1.96\sqrt{\frac{0.297\cdot(1-0.297)}{721}}\ \right]$$

dvs. $[\ 0.264\ ,\ 0.330\ ]$

Fakta om januarmålingen

- Resultatet må tolkes innenfor feilmarginer som varierer mellom 0,9 og 3,3 prosentpoeng.
- Arbeiderpartiet har en feilmargin på 3,3 prosentpoeng, med en øvre grense på 33 prosent og en nedre grense på 26,4 prosent.
- Høyre har en feilmargin 3,2 prosentpoeng, med en øvre grense på 28,1 prosent og en nedre grense på 21,8 prosent.
- 955 personer er blitt spurt og 721 har avgitt partipreferanse.
- Undersøkelsen er gjort mellom 5. og 11. januar.

Faktaene til venstre ble publisert av NRK sammen med meningsmålingen for januar 2016

Vi ser at feilmarginene som oppgis der nettopp er de vi finner ved å bestemme 95% konfidensintervall slik vi gjorde i eksempel 11.2 og oppgave 117a

Eksempel 11.3: En hudlege ønsker å finne ut hvor stor andel av pasienter med psoreasis som vil bli kvitt utslettene hvis de bruker en ny salve

Vi tenker oss at hun lar 150 pasienter som nettopp har fått psoreasis prøve den nye salven, og at 54 av dem blir kvitt utslettene.

Hva kan hun slutte av dette?

Legen er ikke bare interessert i de 150 pasientene. Hun er interessert i hvordan salven vil virke for psoreasispasienter generelt

Det er ikke mulig å trekke et tilfeldig utvalg av alle nåværende og kommende psoreasispasienter

Men hvis det ikke skjer noen endring i pasientgruppen over tid, kan det være rimelig å se på de 150 pasientene som et tilfeldig utvalg av populasjonen av alle nåværende og framtidige pasienter

Under denne forutsetningen får legen følgende estimatet for andelen som blir kvitt utslettene

$\hat{p} = \frac{54}{150} = 0.360$ dvs. 36.0%

For å få en «feilmargin» beregner legen et 95% konfidensintervall:

$$\left[\ 0.360 - 1.96\sqrt{\frac{0.360 \cdot 0.640}{150}}\ ,\ \ 0.360 + 1.96\sqrt{\frac{0.360 \cdot 0.640}{150}}\ \right]$$

Dvs.: $[\ 0.283\ ,\ 0.437\ ]$

Legen kan «regne med» at mellom 28.3% og 43.7% av pasientene vil bli kvitt utslettene hvis de bruker den nye salven

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.

Løsning:

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

Hva betyr det at vi har et 95% konfidensintervall?

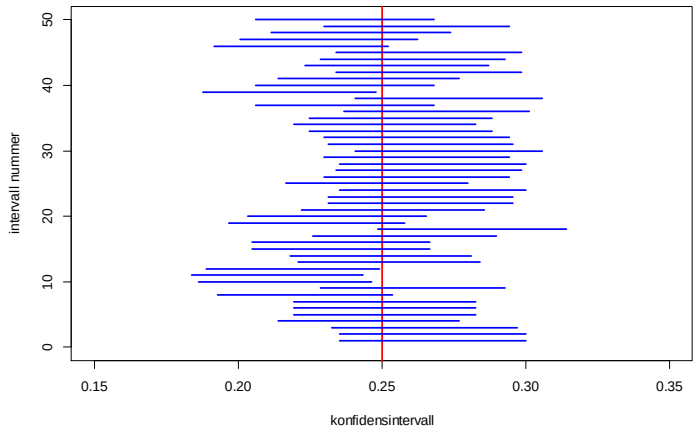
Da vi utledet intervallet, fant vi at det er (tilnærmet) 95% sannsynlig at undersøkelsen vil gi et resultat som er slik at p blir liggende i intervallet

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

Merk at p er et gitt tall mellom 0 og 1 (men ukjent for oss)

Det er grensene i konfidensintervallet som er tilfeldige variabler

Simulering av 50 konfidensintervall (n=721, p=0.25)



Et 95% konfidensintervall vil «i det lange løp» inneholde den sanne verdien av p 95 ut av 100 ganger