MAT0100V
Sannsynlighetsregning og kombinatorikk

Forventning, varians og standardavvik
Tilnærming av binomiske sannsynligheter
Konfidensintervall og hypotesetesting

Ørnulf Borgan
Matematisk institutt
Universitetet i Oslo

Forventningsverdi

Sannsynlighetsfordelingen til en tilfeldig variabel X gir sannsynligheten for de ulike verdiene X kan anta

Vi ønsker i tillegg et summarisk mål som forteller oss hvor fordelingen er «plassert» på tallinja

Forventningsverdien er et slikt summarisk mål

Vi vil bruke rulett som motivasjon (avsnitt 8.1)

Ruletthjulet har 37 felt som er nummerert fra 0 til 36

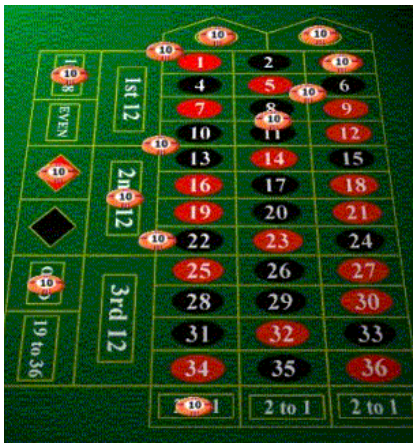
Når ruletthjulet snurrer slippes en liten kule oppi

Kula blir liggende på ett av de 37 nummererte feltene når hjulet stopper

Feltene 1 - 36 er **røde** eller **sorte**, mens 0 er **grønt**

Spillerne setter sin innsats på grupper av felt (det er ikke lov å satse på 0)

Hvis en spiller satser et beløp på k felt og kula stopper på et av dem, vinner spilleren og hun får utbetalt 36/k ganger innsatsen



Vi ser på en «forsiktig» spiller som satser 10 euro på 18 felt (f. eks. de røde)

Spilleren får 20 euro hvis hun vinner og ingenting hvis hun taper. Uansett beholder kasinoet innsatsen på 10 euro

Spillerens nettogevinst i en spilleomgang er 10 euro hvis hun vinner, og den er -10 euro hvis hun taper

Kvinnen spiller tre omganger på denne måten

La Y være hennes samlede nettogevinst i de tre omgangene

Sannsynlighetsfordelingen til Y :

$$P(Y=-30) = \left(\frac{19}{37}\right)^3 = 0.135$$ (taper 3 ganger)

$$P(Y=-10) = 3\cdot\frac{18}{37}\cdot\left(\frac{19}{37}\right)^2 = 0.385$$ (vinner 1 gang og taper 2 ganger)

$$P(Y=10) = 3\cdot\left(\frac{18}{37}\right)^2\cdot\frac{19}{37} = 0.365$$ (vinner 2 ganger og taper 1 gang)

$$P(Y=30) = \left(\frac{18}{37}\right)^3 = 0.115$$ (vinner 3 ganger)

Anta at kvinnen kveld etter kveld spiller tre omganger rulett

Etter N kvelder er hennes gjennomsnittlig nettogevinst:

$$-30\cdot r_N(-30) - 10\cdot r_N(-10) + 10\cdot r_N(10) + 30\cdot r_N(30)$$

Relative frekvenser av de mulige verdiene av nettogevinsten

Når N øker vil gjennomsnittet vil nærme seg

$$-30\cdot P(Y=-30) - 10\cdot P(Y=-10)$$
$$+10\cdot P(Y=10) + 30\cdot P(Y=30) = -0.81$$

Dette er forventningsverdien $E(Y)$

Ruletteksempelet motiverer definisjonen:

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

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

Vi sier ofte forventning i stedet for forventningsverdi

Den greske bokstaven μ («my») brukes for å betegne forventningsverdi

Forventningen er «tyngdepunktet» i fordelingen

Store talls lov

Ruletteksemplet motiverer også store talls lov:

Vi har et forsøk med en tilfeldig variabel X. Hvis vi gjentar forsøket mange ganger, vil gjennomsnittet av verdiene til X nærme seg forventningsverdien E(X)

Store talls lov er blant annet grunnlaget for kasinodrift og forsikringsvirksomhet

To resultater om forventning

Hvis X er binomisk fordelt, er $E(X) = np$

$E(a+bX) = a + b\,E(X)$

Varians

Forventningsverdien til en tilfeldig variabel X forteller oss hva gjennomsnittlig X-verdi vil bli i det lange løp

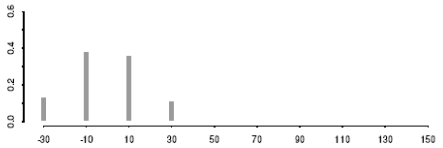
Vi ønsker oss også et summarisk mål som sier noe om hvor mye verdien til en tilfeldig variabel vil variere fra forsøk til forsøk

Variansen er et slikt mål

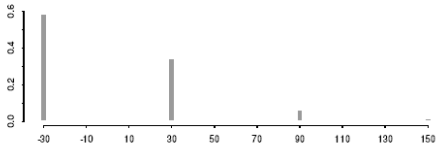
Vi bruker igjen rulett som motivasjon

Vi ser på den «forsiktige» spilleren som tre ganger satser 10 euro på 18 felt og på en annen litt «dristigere» spiller som tre ganger satser 10 euro på 6 felt

Figuren viser fordelingen for nettogevinsten for de to spillerne:



«Forsiktig» spiller (Y)



«Dristig» spiller (Z)

Nettogevinsten Y for den «forsiktige» spilleren og nettogevinsten Z for den «dristige» spilleren har begge forventningsverdi $\mu = -30/37 = -0.81$

Men fordelingen til Z er mer «spredt ut» enn fordelingen til Y

For å få et mål på hvor mye fordelingen til Y er «spredt ut» tar vi utgangspunkt i kvadratavvikene mellom Y-verdiene og forventningsverdien

Hvis Y får verdien -30 er kvadratavviket

$$(-30-\mu)^2 = (-30+30/37)^2 = 852.0$$

Kvinnen spiller kveld etter kveld tre omganger rulett.
Etter N kvelder er det gjennomsnittlige kvadratavviket

$$(-30-\mu)^2 \cdot r_N(-30) + (-10-\mu)^2 \cdot r_N(-10) + (10-\mu)^2 \cdot r_N(10) + (30-\mu)^2 \cdot r_N(30)$$

Relative frekvenser av de mulige verdiene av nettogevinsten

Når N øker, nærmer dette seg

$$(-30-\mu)^2 \cdot P(Y=-30) + (-10-\mu)^2 \cdot P(Y=-10)$$
$$+(10-\mu)^2 \cdot P(Y=10) + (30-\mu)^2 \cdot P(Y=30) \ = 300$$

Denne summen kaller vi variansen til Y
Den skriver vi Var(Y). Altså Var(Y)=300

For den «dristige » spilleren får vi tilsvarende at Var(Z)=1467

Ruletteksempelet motiverer definisjonen:

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

Da er variansen

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

Ofte bruker en σ^2 for å betegne varians

Eksempel 9.1: Vi kaster to terninger, og lar X være summen av antall øyne

k	2	3	4	5	6	7	8	9	10	11	12
$P(X=k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Vi har funnet $E(X) = 7$

Variansen blir:

$$\mathrm{Var}(X) = (2-7)^2 \cdot \frac{1}{36} + (3-7)^2 \cdot \frac{2}{36} + (4-7)^2 \cdot \frac{3}{36}$$
$$+ \ldots + (11-7)^2 \cdot \frac{2}{36} + (12-7)^2 \cdot \frac{1}{36}$$
$$= \frac{210}{36} = \frac{35}{6}$$

Standardavvik

Nettogevinsten til den «forsiktige» spilleren har varians 300

Benevningen for variansen er «kvadrateuro»

Et mål for spredning som har «riktig» benevning er standardavvik:

Standardavviket til en tilfeldig variabel X er gitt ved $SD(X) = \sqrt{\text{Var}(X)}$

Ofte bruker en σ for å betegne standardavvik

Nettogevinsten til den «forsiktige» spilleren har standardavvik 17.30 euro

Varians for binomisk fordeling

Eksempel 9.2: I en søskenflokk er det fire barn

X = «antall gutter i søskenflokken» er binomisk fordelt med n=4 og p=0.514

Har fra før at

$P(X=0)=0.056$ $P(X=1)=0.236$
$P(X=2)=0.374$ $P(X=3)=0.264$ $P(X=4)=0.070$
$\mu = E(X) = 2.06$

Variansen blir

$$\text{Var}(X) = (0-\mu)^2 \cdot 0.056 + (1-\mu)^2 \cdot 0.236 + (2-\mu)^2 \cdot 0.374$$
$$+ (3-\mu)^2 \cdot 0.264 + (4-\mu)^2 \cdot 0.070 = 1.00$$

I eksemplet fant vi (avrundet) Var(X) = 1.00

Merk at $4 \cdot 0.514 \cdot (1 - 0.514) = 1.00$ (avrundet)

Kan vise at vi har generelt:

Hvis X er binomisk fordelt, er Var(X) = $np(1-p)$

Eksempel 9.3. En bestemt type frø spirer med 70% sannsynlighet. Vi sår 20 frø

Variansen til antall frø som spirer er $20 \cdot 0.70 \cdot 0.30 = 4.20$

Variansen til $a + bX$

La X være en tilfeldig variabel med forventningsverdi μ_X

$Y = a + bX$ har forventningsverdi
$\mu_Y = a + b\mu_X$

Kvadratavviket for Y blir

$$(Y-\mu_Y)^2 = (a+bX-\{a+b\mu_X\})^2 = b^2(X-\mu_X)^2$$

Det motiverer resultatet:

Var($a+bX$) = b^2 Var(X)

Eksempel 9.4: Vi ser på den «forsiktige» rulett-spilleren som tre ganger satser 10 euro på 18 felt

La X være antall ganger hun vinner

X er binomisk fordelt med $n = 3$ og $p = 18/37$

$$\mathrm{Var}(X) = 3 \cdot \frac{18}{37} \cdot \frac{19}{37} = 0.749$$

Samlet nettogevinst: $Y = -30 + 20X$

Dermed:

$$\mathrm{Var}(Y) = \mathrm{Var}(-30 + 20X) = 20^2 \mathrm{Var}(X)$$

$$= 400 \cdot 0.749 = 300$$

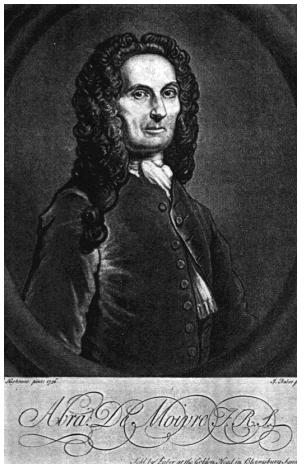
Tilnærming av binomiske sannsynligheter

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

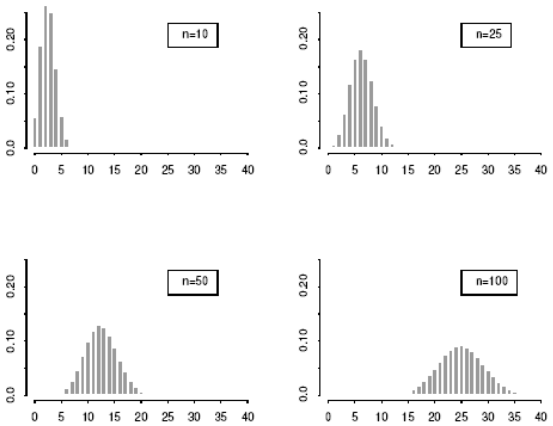
Alt i 1733 viste Abraham de Moivre hvordan en kan finne tilnærmingsverdier for binomiske sannsynligheter

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

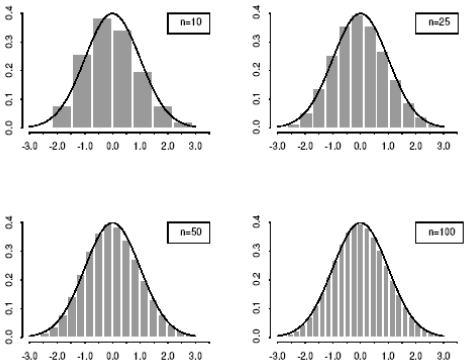
$$Z = \frac{X - E(X)}{SD(X)} = \frac{X - np}{\sqrt{np(1-p)}}$$

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



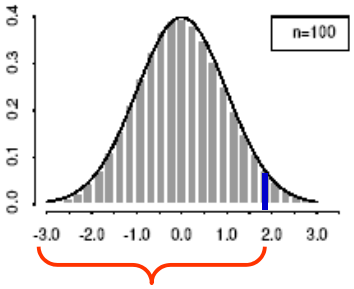
Arealet av en stolpe svarer til sannsynligheten for at Z får den aktuelle verdien

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

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



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:

	Siste desimal i z									
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

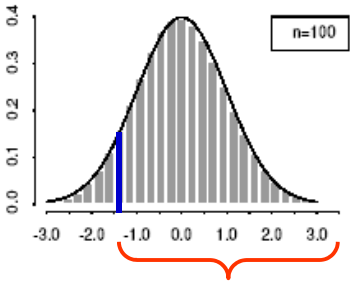
De Moivres tilnærming gir at

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

Vil så bruke de Moivres tilnærming til å finne $P(X \ge 19)$ når n = 100 og p = 0.25

Vi merker oss at

$$P(X \ge 19) = P\left(Z \ge \frac{19 - 100 \cdot 0.25}{\sqrt{100 \cdot 0.25 \cdot 0.75}}\right) = P(Z \ge -1.39)$$



Vi skal egentlig summere arealene av alle søylene til *høyre* for -1.39

Summen av arealene av søylene er omtrent like stor som arealet under $f(x)$ til *høyre* for -1.39

Arealet under standardnormalfordelings-funksjonen til *venstre* for **-1.39** finner vi av tabellen bak i kompendiet:

z	Siste desimal i 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

Arealet til *høyre* for -1.39 er lik én minus arealet til *venstre* for -1.39

Derfor

$$P(X \geq 19) = P(Z \geq -1.39) \approx 1 - 0.082 = 0.918$$

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 $p = 0.32$

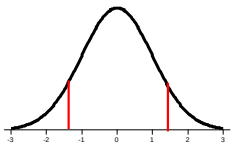
Nå har vi at

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

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

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

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



z	Siste desimal i 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

$$P(300 \le X \le 340) = P\left(-1.36 \le Z \le 1.36\right)$$
$$\approx 0.913 - 0.087 = 0.826$$

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

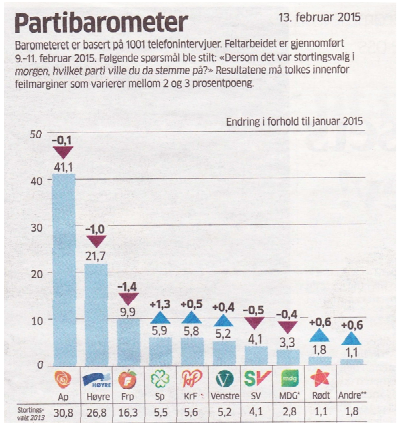
I statistikken gjør vi ikke det. Der er poenget nettopp å kunne si noe om verdien til p når vi har observert X

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 1001 personer som ble intervjuet, ville 411 ha stemt på Arbeiderpartiet hvis det hadde vært valg

Arbeiderpartiets oppslutning er 411/1001=0.411, dvs 41.1%. 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 Ap

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.

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

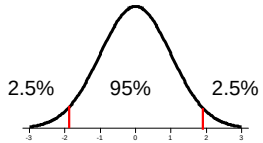
Merk at $\hat{p}$ («p hatt») er en tilfeldig variabel

I eksempelet fikk $\hat{p}$ verdien 0.411

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:

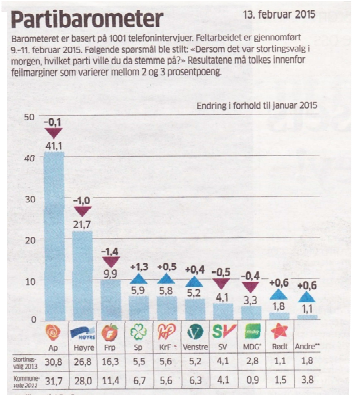
$$P\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 0.95$$

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 Aps oppslutning er

$$\hat{p} = \frac{411}{1001} = 0.411$$

95% konfidens-intervall:

$$\left[\ 0.411 - 1.96\sqrt{\frac{0.411\cdot(1-0.411)}{1001}}\ ,\quad 0.411 + 1.96\sqrt{\frac{0.411\cdot(1-0.411)}{1001}}\ \right]$$

Dvs.: $\left[\ 0.381\ ,\ 0.441\ \right]$ (dette gir en «feilmargin»)