1 Indledning

I denne note vil vi beskrive hvordan man kan angive et godt bud på værdien af successandsynligheden i en binomialfordeling ud fra et forsøg. Endvidere vil vi beskrive hvordan man kan beregne (tilnærmede) konfidensintervaller på baggrund af vores kendskab til χ^2-tests. Vi vil bruge $\binom{n}{k}$ som betegnelse for binomialkoefficienten, hvor k elementer skal vælges fra en mængde med n elementer. Symbolet $\binom{n}{k}$ udtales ofte n vælg k. F.eks. vil man udtale ligningen $\binom{6}{3} = 20$ som „6 vælg 3 er lig tyve“.

2 Estimation af p i binomialfordelinger

Vi starter med et eksempel for bedre at forstå problemstillingen.

Eksempel 1 (Spillet senet)**.** Det gammel-ægyptiske brætspil senet (se Figur 1) slår man ikke med terninger men bruger pinde, som har en rund og en flad side. Umiddelbart er det svært at vurdere, hvad sandsynligheden er for at få henholdsvis „rund“ og „flad“ i et kast med en sådan pind.

For at få en ide om hvad sandsynlighederne mon kan være, kaster vi en af pindene 100 gange og 66 gange lander pinden med den flade side opad. I dette forsøg er andelen af observationer, hvor pinden endt med den flade side opad i $66/100 = 0.66$, så et bud på sandsynligheden for flad kunne være 66 %. Vi ved dog godt, at hvis vi havde slået f.eks. 101 gange i stedet for 100 gange, så ville flad komme til at udgøre en anden andel end 66 %. Spørgsmålet er nu i hvilken forstand $p = 0.66$ er et godt bud på hvad den ukendte sandsynlighed er.

Definition 1. Ved *maksimum likelihood estimatet* for sandsynlighedsparameteren p forstås den værdi af p, som gør sandsynligheden for det observerede størst mulig. Maksimum likelihood estimatet betegnes $\hat{p}$.

Denne definition er måske noget kryptisk formuleret, så lad os se hvordan det fungerer i eksemplet.

Eksempel 2 (Spillet senet, fortsat)**.** Vi har observeret 66 „succes“ i 100 gentagelser af vores eksperiment. Da antallet af succes er binomialfordelt er sandsynligheden for at få 66 succes lig med

$$P(X = 66) = \binom{100}{66} \cdot p^{66} \cdot (1-p)^{34} .$$

Vi vil finde den værdi af $p \in [0;1]$, som maksimerer $P(X = 66)$. Derfor plotter vi $P(X = 66)$ som funktion af p.

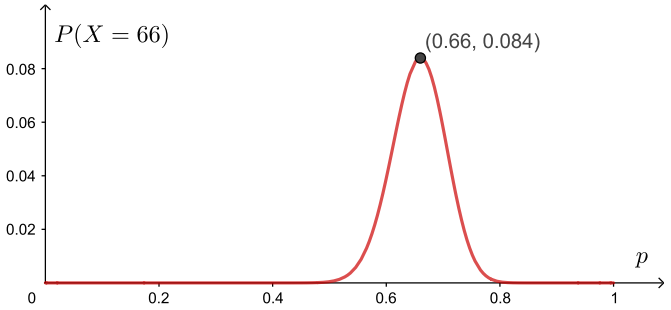
Som det fremgår af Figur 2, er der maksimum for $p = 0.66$. Det kalder vi for maksimum likelihood estimatet for p, hvilket skrives $\hat{p} = 0.66$.

Sætning 1. *Antag at X er binomialfordelt med kendt antalsparameter n og ukendt successandsynlighed p. Hvis man har observeret k succes, så er maksimum likelihood estimatet for p givet ved $\hat{p} = k/n$.*

Bevis. Vi skal bestemme den værdi af p som giver den maksimale værdi af $P(X = k)$. Om sandsynligheden ved vi at $P(X = k) \geq 0$ og $P(X = k) = 0$, hvis $p = 0$ eller hvis $p = 1$. Hvis der kun er et punkt med vandret tangent, så ved vi at der er maksimum i dette punkt. For at gøre de efterfølgende beregninger lidt simplere vil vi maksimere $\ln(P(X = k))$ i stedet for at maksimere $P(X = k)$, idet disse funktioner har maksimum for samme værdi af p . Før vi differntiarer laver vi følgende omskrivning.

$$\begin{aligned}
\ln(P(X = k)) &= \ln\left(\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}\right) \\
&= \ln\binom{n}{k} + \ln\left(p^k\right) + \ln\left((1-p)^{n-k}\right) \\
&= \ln\binom{n}{k} + k \cdot \ln(p) + (n-k) \cdot \ln(1-p) .
\end{aligned}$$

Figur 1: Moderne kopi af det gamle ægyptiske spil senet, hvor man kaster med 4 aflange pinde for at afgøre, hvordan man kan rykke sine brikker. På billedet er 2 pinde landet med den flade (hvide) side opad og 2 er landet med den runde (sorte) side opad.



Figur 2: Plot af likelihood-funktionen.

Denne funktion vil vi nu differentiere, hvilket giver

$$\begin{aligned} 0 + k \cdot \frac{1}{p} + (n-k) \cdot \frac{1}{1-p} \cdot (-1) &= \frac{k}{p} - \frac{n-k}{1-p} \\ &= \frac{k \cdot (1-p)}{p \cdot (1-p)} - \frac{p \cdot (n-k)}{p \cdot (1-p)} \\ &= \frac{k \cdot (1-p) - p \cdot (n-k)}{p \cdot (1-p)} \\ &= \frac{k - k \cdot p - p \cdot n + p \cdot k}{p \cdot (1-p)} \\ &= \frac{k - p \cdot n}{p \cdot (1-p)} \,. \end{aligned}$$

Denne brøk sættes lig nul for at bestemme den værdi af p, hvor tangenten er vandret.

$$\begin{aligned} 0 &= \frac{k - pn}{p \cdot (1-p)} \\ 0 &= k - pn \\ p \cdot n &= k \\ p &= \frac{k}{n} \,. \end{aligned}$$

Derfor er $\hat{p} = {}^{k}/_{n}$. □

3 Konfideninterval

Vi fortsætter med vores eksempel.

Eksempel 3 (Spillet senet, fortsat)**.** Vi har set at maksimum likelihood estimatet er $\hat{p} = 0.66$, men det er jo langt fra sikkert, at dette er den rigtige værdi af p. Måske er den rigtige værdi $p = {}^{1}/_{2}$. Denne hypotese vil vi nu teste med en χ^2-test på et 5 % signifikansniveau.
$H_0 : p = {}^{1}/_{2}$.
$H_a : P \neq {}^{1}/_{2}$.

Udfald	Obs.	Forv.
flad	66	50
rund	34	50
i alt	100	100

Vi udregner χ^2-teststørrelsen.

$$\begin{aligned} \chi^2 &= \frac{(66-50)^2}{100} + \frac{(34-50)^2}{100} \\ &= \frac{256}{100} + \frac{256}{100} \\ &= \frac{512}{100} \\ &= 5.12 \,. \end{aligned}$$

Da den kritiske værdi er $\chi^2_{krit} = 3.84$, ligger den observerede værdi af χ^2-teststørrelsen over den kritiske værdi, og vi vil derfor forkaste nulhypotesen. Konklusionen er, at p ikke er lig med 0.5 . Man kan sige, at den observerede værdi 66 afviger signifikant fra den forventede værdi 50. En anden måde at formulere det samme på er, at hypotesen $p = 0.5$ ligger for langt fra den estimerede værdi $\hat{p} = 0.66$ til at vi kan acceptere nulhypotesen. Rundt om den estimerede værdi $\hat{p} = 0.66$ vil vi nu danne et interval af hypoteser, som kan accepteres.

Vi vil nu lave en tilnærmet formel for intervallet af acceptable hypoteser. Lad k betegne det observerede antal succes i n gentagelser af et eksperiment. Vi betegner maksimum likelihood estimatet k/n med $\hat{p}$.

Udfald	Obs.	Forv.
Succes	k	$n \cdot p$
Fiasko	$n-k$	$n \cdot (1-p)$
I alt	n	n

Vi udregner χ^2-teststørrelsen.

$$\begin{aligned}
\chi^2 &= \frac{(k - n\cdot p)^2}{n\cdot p} + \frac{((n-k) - n\cdot(1-p))^2}{n\cdot(1-p)} \\
&= \frac{(k - n\cdot p)^2}{n\cdot p} + \frac{(n-k-n+n\cdot p)^2}{n\cdot(1-p)} \\
&= \frac{(k - n\cdot p)^2}{n\cdot p} + \frac{(-k+n\cdot p)^2}{n\cdot(1-p)} \\
&= (k - n\cdot p)^2 \cdot \left(\frac{1}{n\cdot p} + \frac{1}{n\cdot(1-p)}\right) \\
&= n^2\left(\frac{k}{n} - p\right)^2 \cdot \left(\frac{1-p}{n\cdot p\cdot(1-p)} + \frac{p}{n\cdot p\cdot(1-p)}\right) \\
&= n^2(\hat{p}-p)^2 \cdot \frac{1}{n\cdot p\cdot(1-p)} \\
&= \frac{n\cdot(\hat{p}-p)^2}{p\cdot(1-p)}\,.
\end{aligned}$$

Vi vil acceptere vores nulhypotese, hvis teststørrelsen ikke overstiger den kritiske værdi - altså

$$\frac{n\cdot(\hat{p}-p)^2}{p\cdot(1-p)} \le \chi^2_{krit}\,.$$

For at gøre de efterfølgende udregninger lidt lettere vil vi tilnærme p i nævneren med $\hat{p}$, hvorved vi får

$$\frac{n\cdot(\hat{p}-p)^2}{p\cdot(1-p)} \approx \frac{n\cdot(\hat{p}-p)^2}{\hat{p}\cdot(1-\hat{p})}\,. \tag{1}$$

Vi vil derfor acceptere vores nulhypotese, hvis

$$\begin{aligned}
\frac{n\cdot(\hat{p}-p)^2}{\hat{p}\cdot(1-\hat{p})} &\le \chi^2_{krit} \\
(\hat{p}-p)^2 &\le \chi^2_{krit}\frac{\hat{p}\cdot(1-\hat{p})}{n} \\
|p-\hat{p}| &\le \left(\chi^2_{krit}\frac{\hat{p}\cdot(1-\hat{p})}{n}\right)^{1/2} \\
|p-\hat{p}| &\le \chi_{krit}\cdot\left(\frac{\hat{p}\cdot(1-\hat{p})}{n}\right)^{1/2}\,.
\end{aligned}$$

Her er χ_{krit} kvadratroden af χ^2_{krit}.

α	$1-\alpha$	χ^2_{krit}	χ_{krit}
10 %	90 %	2.7055	1.6449
5 %	95 %	3.8415	1.9600
1 %	99 %	6.6349	2.5758

Man kan med andre ord acceptere enhver nulhypotese, hvor p ligger i intervallet med endepunkter

$$\hat{p} \pm \chi_{krit} \cdot \left(\frac{\hat{p} \cdot (1-\hat{p})}{n} \right)^{1/2} . \tag{2}$$

Hvis signifikansniveauet er α, vil vi sige at formlen beregner et $1-\alpha$ *konfidensinterval*. Hvis vi f.eks. har $\alpha = 5\,\%$, så giver formlen et 95 % konfidensinterval.

Eksempel 4 (Spillet senet, fortsat)**.** I vores eksempel er $\hat{p} = 0.66$ og $\chi_{krit} = 1.96$, så intervallet af acceptable hypoteser er

$$\begin{aligned} 0.66 \pm 1.96 \left(\frac{0.66 \cdot (1-0.66)}{100} \right)^{0.5} &= 0.66 \pm 0.093 \\ &= \begin{cases} 0.753 \\ 0.567 \end{cases} \end{aligned}$$

Intervallet $[0.57; 0.75]$ kaldes et 95 % konfidensinterval, og man vil acceptere nulhypoteser om at p har en bestemt værdi, hvis værdien ligger i dette interval. Egentlig burde vi have løst ligningen

$$\begin{aligned} \frac{n \cdot (\hat{p}-p)^2}{p \cdot (1-p)} &= \chi^2_{krit} \\ \frac{100 \cdot (0.66-p)^2}{p \cdot (1-p)} &= 3.8415 \\ 100 \cdot (0.66-p)^2 &= 3.8415 \cdot p \cdot (1-p) \end{aligned}$$

Denne 2.-gradsligning har løsningerne $p = 0.563$ og $p = 0.745$. Som det ses, har vi lavet en mindre fejl, da vi erstattede p med $\hat{p}$ i ligning (1). I de fleste praktiske sammenhænge er fejlen så ubetydelig, at vi vil se bort fra den. Det betyder dog, at værdierne af endepunkterne for intervallet er tilnærmede værdier og derfor ikke bør angives med for mange decimaler.

Tommelfingerregel Hvis $10 < n \cdot p < n - 10$ for alle p i intervallet givet ved formlen (2), vil formlen med god tilnærmelse give konfidensintervaller for ukendt andel i binomialfordelinger.