

Konfidensinterval

Det er ofte tilfældet at man har brug for at beregne en sandsynlighed ud fra en stikprøve. Det kunne f.eks. være at man ville beregne sandsynligheden for at slå en sekser med en skæv terning hvor man ikke på forhånd kan bestemme sandsynligheden på andre måder og man derfor må foretage en masse terningkast, dvs. en stikprøve, for at beregne hvor mange seksere der forekommer i stikprøven. Men hvor godt kan man forvente at denne sandsynlighed der er beregnet ud fra en tilfældig stikprøve passer med den korrekte sandsynlighed? Til at håndtere denne problemstilling anvendes begrebet *konfidensinterval* hvor man beregner et interval omkring den beregnede sandsynlighed hvor den sande værdi med 95% sandsynlighed ligger inden for.

I tilfælde af en binomial proces hvor betingelserne er opfyldt for at sandsynlighedsfordelingen kan approksimeres med en normalfordeling kan man netop anvende egenskaberne fra normalfordelingen til at bestemme et konfidensinterval.

Hvis man forestiller sig at man foretager en stikprøve på n værdier hvoraf r er antallet af succeser kan man beregne frekvensen f af succeser som en approksimeret sandsynlighed $\hat{p}$ for succes:

p-hat = f = r/n

Ud fra denne approksimerede sandsynlighed kan man beregne den approksimerede middelværdi $\hat{\mu}$ og spredning $\hat{\sigma}$:

mu-hat = n * p-hat og sigma-hat = sqrt(n * p-hat * (1 - p-hat))

Der antages, at normalfordelingen kan anvendes som approksimation til binomialfordelingen. Dette vil sige, at et tilfældigt udfald med 95.45% sandsynlighed vil ligge inden for intervallet der dækker 2 spredninger på hver side af middelværdien. Idet den sande middelværdi af binomialprocessen i sig selv kan betragtes som et udfald, og faktisk det forventede udfald, vil det gælde at:

P(mu-hat - 2*sigma-hat <= mu <= mu-hat + 2*sigma-hat) = 95.45%

Man kan altså med lidt over 95% sandsynlighed sige at:

mu-hat - 2*sigma-hat <= mu <= mu-hat + 2*sigma-hat

For at opnå et interval for selve sandsynligheden p omskrives middelværdien med $n \cdot p$ og den approksimerede middelværdi samt spredning omskrives ved de foregående udtryk med den approksimerede sandsynlighed $\hat{p}$. Herved opnås:

n * p-hat - 2*sqrt(n * p-hat * (1 - p-hat)) <= n * p <= n * p-hat + 2*sqrt(n * p-hat * (1 - p-hat))

For nu at opnå et interval for sandsynligheden p divideres med n :

p-hat - 2*sqrt(p-hat * (1 - p-hat) / n) <= p <= p-hat + 2*sqrt(p-hat * (1 - p-hat) / n)

Resultatet kan nu opsummeres i følgende sætning.

Konfidensinterval

For en binomialfordelt stikprøve, hvor det antages at normalfordelingsapproksimationen gælder, på n værdier hvoraf r er antallet af succeser kan sandsynligheden for en succes approksimeres ved:

p-hat = r/n

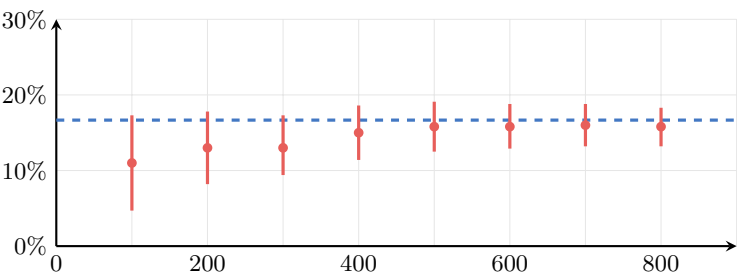
Ud fra denne approksimerede sandsynlighed kan man med 95.45% sandsynlighed sige at den sande værdi p for sandsynligheden for en succes ligger inden for følgende konfidensinterval:

p in [p-hat - 2*sqrt(p-hat * (1 - p-hat) / n), p-hat + 2*sqrt(p-hat * (1 - p-hat) / n)]

For at illustrere anvendelsen af konfidensintervaller er der udført et forsøg med en 6-sidet terning for ved hjælp af en stikprøve at bestemme sandsynligheden for at slå en sekser. Terningen blev kastet 800 gange og for hvert 100 kast noteres antallet af seksere som ses i tabellen nedenfor.

n	100	200	300	400	500	600	700	800
r	11	26	40	60	79	95	112	126

På baggrund af resultaterne beregnes for hvert 100 kast den approksimerede sandsynlighed samt 95% konfidensintervallet. Resultatet er vist grafisk i figuren nedenfor sammen med den sande værdi for sandsynlighed på 16.67% (stiplet blå linje).



Det ses af figuren, at for hver approksimation af sandsynligheden ligger den sande værdi for sandsynlighed inden for konfidensintervallet. Ydermere ses at jo større stikprøven bliver jo mindre bliver konfidensintervallet hvilket vil sige at den approksimerede sandsynlighed vil ligge tættere på den sande værdi af sandsynlighed jo større stikprøven bliver.