Spørsmål fra salen:

Ang betydningen av konfidensintervall. Boka sier at det IKKE betyr at det feks er 95 % sannsynlighet for at den ekte my havner innenfor intervallet, men at det betyr at dette intervallet er kalkulert ved en metode som gir den ekte my i 95% av alle mulige utvalg. (boka s.366)

Så det blir altså helt feil å tenke på det som sannsynlighet?

Svar:

Dette er veldig observant og en veldig god kommentar. Boka har rett. Din mageølelse når du tenker på det som sannsynlighet er også nesten rett. Men grunnen til at det bare er nesten rett, har med det stokastiske å gjøre:

My er et tall, altså ett enkelt tall, selv om det er ukjent. Mao: Det er ikke en stokastisk variabel, og har ikke en sannsynlighet knyttet til seg. My er my, bare ukjent.

Det betyr at når deu først har regnet ut et konfidensintervall, vil den ukjente my-en enten være i intervallet, eller ikke være i intervallet. Vi vet ikke hva som er sant, men det er ingen sannsynlighet knyttet til et ferdig utregnet intervall. Intervallet er bare to tall som vi har beregnet og vet hva er. My vet vi ikke hva er, men den er altså ikke stokastisk, bare ukjent. Altså har vi tre faste tall: Nedre og øvre grense i konfidensintervallet, og my, som enten er mellom disse verdiene eller ikke. Det er ingen sannsynlighet (stokastikk) knyttet tl noen av dem.

Hvis du skal tenke på et konfidensintervall som sannsynlighet, går det an å gjøre det når du snakker om det u-utregnede konfidensintervallet, altså når vi skriver
$P(\text{xstrek}-1.96*SE(\text{xstrek}) < my < \text{xstrek}+1.96*SE(\text{xstrek})) = 0.95$

Så lenge vi uttaler oss om xstrek (altså gjennomsnittet) som en stokastisk variabel som ennå ikke er observert eller regnet ut, vil dette uttrykket gi mening som sannsynlighet. Selv om det altså er nedre og øvre grense som er stokastisk, og ikke my. (Vi er vant til å regne ut sannsynligheter av typen $P(a < Z < b)=0.95$, når Z er en stokastisk variabel, men her er det yttergrensene i ulikheten som er det stokastiske, ikke midten.)

Så hvis det skal være formelt riktig å tenke sannsynlighet i forbindelse med konfidensintervall, er det *grensene* i intervallet som er utgangspunktet. Altså kan vi ikke si "sannsynligheten for at my er i intervallet er 95%". Det er riktigere da å si "sannsynligheten for at intervallet dekker my er 95%", men det er heller ikke riktig når intervallet er ferdig utregnet.

Det vi derimot kan si, og som er helt riktig, er at "95% av alle utregnede 95% konfidensintervaller inneholder my, og 5% av alle utregnede 95% KI gjør det ikke".

Unntaket er i Bayesiansk statistikk, som ikke er pensum i STK1000. I Bayesiansk statistikk ser man på parameteren my som en stokastisk størrelse. Da har my en (apriori) sannsynlighetsfordeling for my som oppdateres med kunnskap fra observerte data, og vi kan regne ut det såkalte 95% kredibilitetsintervallet for my, KrI.

Et KrI er svært likt et KI, men siden my nå er en stokastisk størrelse, kan vi med god samvittighet si "det er 95% sannsynlig at my er i det utregnede kredibilitetsintervallet", og det er riktig.

Så grunnen til at mageølelsen din om KI er ganske riktig, men formelt feil, er altså hvor det stokastiske er definert.

Forhåpentligvis var dette oppklarende for flere.