Hvorfor *n*-1 i stikprøvevariansen?

©iStock.com/choness

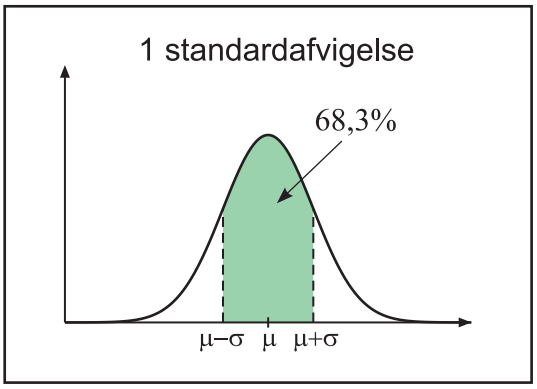
Lad os sige, at en fabrik producerer en bestemt type halogenpærer. Det viser sig, at levetiden for en sådan elpære varierer efter en *normalfordeling*. Nogle elpærer vil ikke holde så længe som andre vil. Det skyldes tilfældigheder i brugen og i produktionen. En normalfordeling er specificeret gennem to parametre: *middelværdien* μ og spredningen σ. Lad os i første omgang forestille os, at vi vidste, at middelværdien var 2000 timer og spredningen var 420 timer (fiktive værdier). På grund af egenskaberne for normalfordelingen – som dog ikke skal gennemgås på dette sted – kan man vise, at sandsynligheden for at levetiden for en tilfældig valgt elpære af nævnte type højst falder spredningen til hver side i forhold til middelværdien, er ca. 68%. På tilsvarende vis kan det vises, at sandsynligheden for at elpæren har en levetid, som højst falder 2 gange spredningen til hver side i forhold til middelværdien, er ca. 95%. Det kan skrives matematisk således:

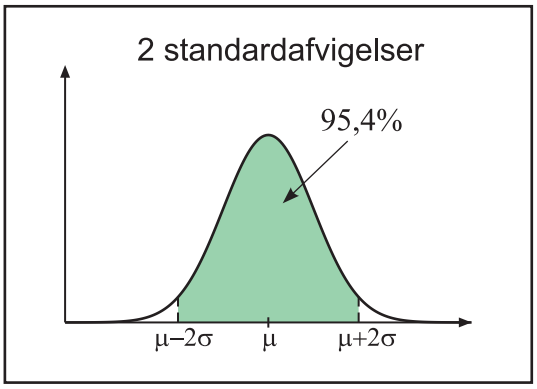
$$P(\mu-\sigma\le X\le\mu+\sigma)\approx 68\% \quad \text{og} \quad P(\mu-2\sigma\le X\le\mu+2\sigma)\approx 95\%$$

eller med talværdier indsat:

$$P(2000-420\le X\le 2000+420) \;=\; P(1580\le X\le 2420) \;\approx\; 68\%$$
$$P(2000-2\cdot 420\le X\le 2000+2\cdot 420) \;=\; P(1160\le X\le 2840) \;\approx\; 95\%$$

hvor P læses "sandsynligheden for" og X er den stokastiske variabel, som angiver levetiden for den udtrukne elpære. Spredningen fortæller altså ganske meget i tilfældet med en normalfordeling. 68% af den nævnte type elpærer vil altså have en levetid på mellem 1580 og 2420 timer. Spredningen kaldes undertiden også for *standardafvigelsen*.





Nu antager vi imidlertid, at vi *ikke* kender parametrene μ og σ. Altså en helt ny situation. Vi kan da udtrække en tilfældig stikprøve fra produktionen med henblik på at bestemme et *estimat* af de to parametre. En stikprøve med størrelsen 25 gav følgende resultater:

1720, 1740, 2303, 1947, 2166, 2237, 2318, 1403, 2546, 2636, 1859, 2611, 1713, 2244, 1711, 2197, 2605, 2228, 1938, 1382, 1845, 1326, 2403, 1240, 1984. (enhed: timer)

Det er sikkert ikke overraskende for de fleste, at et godt bud på middellevetiden μ fås ved at tage den *empiriske middelværdi* af stikprøveværdierne, som vi betegner $x_1, x_2, \ldots, x_n$:

$$\text{(1)} \qquad \overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

I dette tilfælde vil det give følgende resultat:

$$\begin{aligned} \overline{x} &= \frac{1}{25} \cdot (1720 + 1740 + 2303 + 1947 + 2166 + 2237 + 2318 + 1403 + 2546 \\ &\quad + 2636 + 1859 + 2611 + 1713 + 2244 + 1711 + 2197 + 2605 + 2228 \\ &\quad + 1938 + 1382 + 1845 + 1326 + 2403 + 1240 + 1984) \\ &= 2012{,}0 \end{aligned}$$

Den *empiriske varians* s^2 fås af følgende formel:

$$\text{(2)} \qquad s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad \text{(Stikprøvevariansen)}$$

Den *empiriske spredning* eller *standardafvigelsen* s (Engelsk: *Standard Deviation*) fås herefter ved at tage kvadratroden:

$$\text{(3)} \qquad s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2} \quad \text{(Stikprøvespredningen)}$$

I dette tilfælde fås følgende værdi for stikprøvevariansen:

$$\begin{aligned} s^2 &= \frac{1}{25-1} \cdot \Big[(1720 - 2012{,}0)^2 + (1740 - 2012{,}0)^2 + (2303 - 2012{,}0)^2 + (1947 - 2012{,}0)^2 \\ &\quad + (2166 - 2012{,}0)^2 + (2237 - 2012{,}0)^2 + (2318 - 2012{,}0)^2 + (1403 - 2012{,}0)^2 \\ &\quad + (2546 - 2012{,}0)^2 + (2636 - 2012{,}0)^2 + (1859 - 2012{,}0)^2 + (2611 - 2012{,}0)^2 \\ &\quad + (1713 - 2012{,}0)^2 + (2244 - 2012{,}0)^2 + (1711 - 2012{,}0)^2 + (2197 - 2012{,}0)^2 \\ &\quad + (2605 - 2012{,}0)^2 + (2228 - 2012{,}0)^2 + (1938 - 2012{,}0)^2 + (1382 - 2012{,}0)^2 \\ &\quad + (1845 - 2012{,}0)^2 + (1326 - 2012{,}0)^2 + (2403 - 2012{,}0)^2 + (1240 - 2012{,}0)^2 \\ &\quad + (1984 - 2012{,}0)^2 \Big] \\ &= 173169 \end{aligned}$$

og dermed stikprøvespredningen: $s = \sqrt{173169} = 416{,}4$.

Forskellen mellem parret μ og σ^2 og parret $\overline{x}$ og s^2 er, at det første par er de korrekte værdier for middelværdi og varians, mens det sidste par er *estimater* for disse værdier. Sidstnævnte afhænger således af den udtrukne stikprøve. For en given konfidensgrad $1-\alpha$, kan man angive et *konfidensinterval* omkring $\overline{x}$ således, at sandsynligheden for, at den rigtige værdi for middelværdien (dvs. μ) er i dette interval, er lig med $1-\alpha$. Noget tilsvarende kan gøres omkring s^2. Det involverer dog students t-fordelingen og χ^2 -fordelingen, og vil være alt for omfattende at komme ind på her.

Der er en ting, som stikker i øjnene, og det er udseendet af stikprøvevariansen i (2). Det er egentlig hovedformålet med dette tillæg at redegøre for det. Umiddelbart ville det være mere naturligt at definere den med n i nævneren i stedet for $n-1$. For at forstå, hvorfor det sidste er mest fornuftigt, er vi nødt til at tale lidt om begrebet en *estimator*.

En estimator uden skævhed

En *estimator* er en regel eller metode til at opnå et *estimat* for værdien af en størrelse eller parameter, baseret på observeret data. Parameteren kan for eksempel være middelværdien eller variansen i en fordeling.

I det følgende vil jeg antage, at læseren er bekendt med basal teori om stokastiske variable. Lad der være givet n uafhængige stokastiske variable $X_1, X_2, \ldots, X_n$ hver med den samme fordeling med middelværdi μ og varians σ^2. Man kan tænke på X_i som den stokastiske variabel, som angiver udfaldet af den i'te "udtrækning" i en stikprøve. I tilfældet med elpærerne ville X_i for eksempel kunne angive levetiden for den i'te elpære i stikprøven. En estimator for middelværdien μ er oplagt følgende:

$$\text{(4)} \qquad \overline{X} = \frac{1}{n}\cdot\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Det er en regel eller formel, som givet en stikprøve, kan give et estimat for middelværdien μ. Estimatet fås naturligvis ved at indsætte værdierne for de stokastiske variable, som vi fik fra stikprøven. Vi kan passende kalde værdierne $x_1, x_2, \ldots, x_n$. I eksemplet med elpærerne ville det være de 25 levetider. På den måde giver den empiriske middelværdi $\overline{x}$ fra (1) et estimat for middelværdien baseret på stikprøveresultatet. Estimatoren er samtidigt en stokastisk variabel, hvilket sætter os i stand til at udregne middelværdi og varians for denne via de almindelige definitioner for stokastiske variable.

Sætning 1

Lad der være givet n uafhængige stokastiske variable $X_1, X_2, \ldots, X_n$ hver med den samme fordeling med middelværdi μ og varians σ^2. Da gælder følgende:

a) $\mu_{\overline{X}} = E(\overline{X}) = \mu$

b) $\sigma^2_{\overline{X}} = Var(\overline{X}) = \dfrac{\sigma^2}{n}$

Bevis: Vi udnytter lineariteten af middelværdien også kaldet den forventede værdi:

$$\begin{aligned} E(\overline{X}) &= E\left(\tfrac{1}{n}\cdot X_1 + \tfrac{1}{n}\cdot X_2 + \ldots + \tfrac{1}{n}\cdot X_n\right) \\ &= \tfrac{1}{n}\cdot E(X_1) + \tfrac{1}{n}\cdot E(X_2) + \ldots + \tfrac{1}{n}\cdot E(X_n) \\ &= \tfrac{1}{n}\cdot\mu + \tfrac{1}{n}\cdot\mu + \ldots + \tfrac{1}{n}\cdot\mu \\ &= \mu \end{aligned}$$

Ved hjælp af regnereglen for variansen af en linearkombination af *uafhængige* stokastiske variable fås følgende resultat:

$$\begin{aligned} Var(\overline{X}) &= Var\left(\tfrac{1}{n}\cdot X_1+\tfrac{1}{n}\cdot X_2+\ldots+\tfrac{1}{n}\cdot X_n\right) \\ &= \left(\tfrac{1}{n}\right)^2\cdot Var(X_1)+\left(\tfrac{1}{n}\right)^2\cdot Var(X_2)+\ldots+\left(\tfrac{1}{n}\right)^2\cdot Var(X_n) \\ &= \left(\tfrac{1}{n}\right)^2\cdot\sigma^2+\left(\tfrac{1}{n}\right)^2\cdot\sigma^2+\ldots+\left(\tfrac{1}{n}\right)^2\cdot\sigma^2 \\ &= n\cdot\left(\frac{1}{n}\right)^2\cdot\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

□

Egenskaben a) udtaler, at *middelværdien* af estimatoren er lig med den værdi, som estimatoren skal estimere (her μ). Har en estimator en sådan egenskab, kaldes den *central*. En estimator, som ikke er central, siges at have en *skævhed* eller *bias*. Man skal tænke på det på denne måde: Den formel, som estimatoren giver anledning til, skal jo bruges til at indsætte værdier fra en stikprøve. Hver stikprøve vil normalt give et forskelligt estimat, men her er det altså rart, at man teoretisk kan stole på, at stikprøver i gennemsnit vil give den værdi, der skal estimeres.

Egenskaben b) i sætning 1 er også interessant. Den fortæller nemlig, at variansen og dermed spredningen for et gennemsnit er mindre end spredningen på hver af størrelserne. Det forklarer også, hvorfor det er en god idé i fysik og andre naturvidenskabelige fag at tage flere målinger af den samme størrelse og til sidst tage gennemsnittet. Så bliver usikkerheden nemlig mindre!

Vi er nu klar til at kigge på en estimator for variansen. Som bekendt er variansen for en diskret stokastisk variabel med middelværdi μ generelt set givet ved et udtryk på formen:

$$Var(X) = \sum_{i=1}^{N}(x_i-\mu)^2\cdot P(X=x_i)$$

Denne giver os idéen til en oplagt kandidat til en estimator for variansen, idet vi udskifter $P(X=x_i)$ med vægten $1/n$, udskifter den ukendte middelværdi μ med stikprøvegennemsnittet $\overline{X}$ og ellers summerer over alle elementer i stikprøven:

$$(5)\qquad b^2=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\frac{\left(X_1-\overline{X}\right)^2+\left(X_2-\overline{X}\right)^2+\ldots+\left(X_n-\overline{X}\right)^2}{n}$$

Det skal lige tilføjes, at såfremt de stokastiske variable X_i er uafhængige normalfordelte stokastiske variable med middelværdi μ og varians σ^2, så dukker (5) faktisk op som en såkaldt Maximum Likelihood estimator for variansen, men det er en helt anden sag, som vi ikke skal komme nærmere ind på her. Vi kan roligt tage (5) som et kvalificeret gæt. Vi skal undersøge, om den nye estimator er central eller ej.

Sætning 2

Varians-estimatoren b^2 givet ved (5) har bias. Nærmere bestemt gælder:

(6) $$E(b^2)=\frac{n-1}{n}\cdot\sigma^2$$

Bevis: I det følgende skal vi gøre brug af resultaterne i sætning 1 samt følgende generelle formel for variansen af en stokastisk variabel X:

(7) $$Var(X)=E(X^2)-(E(X))^2 \quad\Leftrightarrow\quad E(X^2)=Var(X)+(E(X))^2$$

Lad os regne på middelværdien af estimatoren b^2, idet vi gør heftig brug af lineariteten af middelværdien (den forventede værdi):

$$\begin{aligned}
E(b^2) &= E\left(\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\\
&= \frac{1}{n}\cdot E\left(\sum_{i=1}^{n}\left(X_i^2-2\cdot X_i\cdot\bar{X}+\bar{X}^2\right)\right)\\
&= \frac{1}{n}\cdot E\left(\sum_{i=1}^{n}X_i^2-2\bar{X}\cdot\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\bar{X}^2\right)\\
&= \frac{1}{n}\cdot E\left(\left(\sum_{i=1}^{n}X_i^2\right)-2\bar{X}\cdot n\cdot\bar{X}+n\cdot\bar{X}^2\right)\\
&= \frac{1}{n}\cdot E\left(\left(\sum_{i=1}^{n}X_i^2\right)-n\cdot\bar{X}^2\right)\\
&= \frac{1}{n}\cdot\left[\left(\sum_{i=1}^{n}E(X_i^2)\right)-n\cdot E(\bar{X}^2)\right]\\
&= \frac{1}{n}\cdot\left[\left(\sum_{i=1}^{n}(\sigma^2+\mu^2)\right)-n\cdot\left(\frac{\sigma^2}{n}+\mu^2\right)\right]\\
&= \frac{1}{n}\cdot\left[n\cdot\left(\sigma^2+\mu^2\right)-n\cdot\left(\frac{\sigma^2}{n}+\mu^2\right)\right]\\
&= \frac{n-1}{n}\cdot\sigma^2
\end{aligned}$$

hvor vi i tredjesidste lighedstegn har benyttet den sidste identitet i (7) for både den stokastiske variabel X_i og den stokastiske variabel $\bar{X}$. Foruden det er sætning 1 brugt.

□

Sætning 2 fortæller os, at hvis man bruger formlen

(8) $$\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2$$

til at bestemme et estimat for variansen, så vil den i middel skyde lidt under den rigtige varians, hvis man foretager mange stikprøver og udregner variansen med formlen. Det er ikke optimalt. Heldigvis kan man hurtigt rette op på det, som følgende sætning viser:

Sætning 3

Estimatoren S^2, givet ved udtrykket nedenfor, er en central estimator for variansen:

(9) $$S^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Bevis: Vi benytter blot lineariteten af middelværdien:

$$E(S^2) = E\left(\frac{n}{n-1}\cdot b^2\right) = \frac{n}{n-1}\cdot E(b^2) = \frac{n}{n-1}\cdot\frac{n-1}{n}\cdot\sigma^2 = \sigma^2$$

□

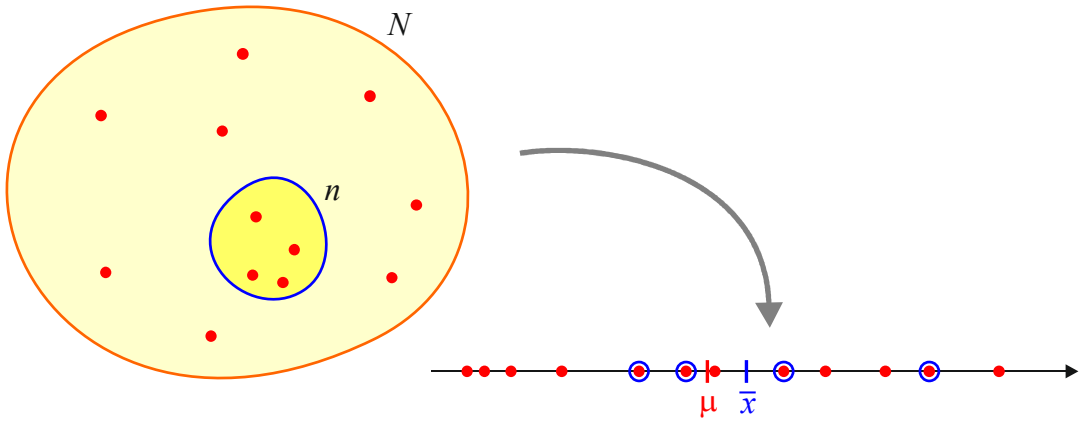
Dermed er det antydet, at formel (2) for stikprøvevariansen er fornuftig. Man kan også studere andre egenskaber for estimatorer, men det vil vi ikke gøre her. En anden ting er, at forskellen på formlen (2) og (8) ikke er stor, hvis stikprøvestørrelsen *n* er stor. Som et estimat for spredningen benyttes kvadratroden af stikprøvevariansen fra (2).

Diverse kommentarer

Pointe 1: En anden ting, som man også kan filosofere over, er følgende: Hvis stikprøven kun indeholdt én måling, så kan formel (8) bruges. Men hvordan i alverden vil en stikprøve på 1 måling overhovedet kunne bruges til at vurdere spredningen i den oprindelige fordeling? Det giver ikke mening. Formel (2) kan derimod ikke bruges for en stikprøve bestående af kun én måling, fordi nævneren da vil give 0! Så formlen for stikprøvevariansen er slet ikke defineret for en stikprøve, som kun består af en måling. Derfor giver (2) meget mere mening.

Pointe 2: Et andet argument, som man også ofte ser fremført er, at man skal dividere med $n-1$ fordi antallet af *frihedsgrader* i udtrykket (2) kun er $n-1$. Før vi overhovedet begynder at estimere variansen, har vi nemlig estimeret middelværdien μ ved hjælp af stikprøvegennemsnittet $\bar{x}$, og det koster 1 frihedsgrad! Dermed er der kun $n-1$ frihedsgrader tilbage til at estimere variansen. Dette er ikke svært at forstå, for kender man værdierne for $x_1, x_2, \ldots, x_{n-1}$, så giver værdien af x_n nemlig sig selv, fordi man kender gennemsnittet: $x_n = n\cdot\bar{x} - (x_1 + x_2 + \ldots + x_{n-1})$. Der er altså kun $n-1$ af x_i'erne, der kan varieres frit. Mens argumentet med antal frihedsgrader er klar, står det knapt så klart, hvorfor man derefter skal dividere med antal frihedsgrader for at få formlen for stikprøvevariansen …

Pointe 3: Lad os gøre situationen med *stikprøve* kontra *population* endnu mere tydelig. Vi antager, at vores population indeholder N talværdier. På figuren er de desuden anbragt på en tallinje. Vi udtrækker ud fra populationen en stikprøve bestående af n elementer. De er også markeret på tallinjen med en blå ring omkring. På tallinjen er også anbragt den rigtige middelværdi μ for populationen. Den kan vi ikke bestemme med stikprøven. Det bedste estimat, vi kan finde for den, er gennemsnittet af stikprøvens værdier $\overline{x}$, også kaldet den empiriske middelværdi. Den er markeret med blåt på tallinjen. Forskellen på de to kan undertiden være stor, afhængig af valgt af stikprøve.



Vi ønsker at estimere den eksakte varians, som er:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} \quad \text{(populationsvariansen)}$$

Man kunne måske i første omgang tænke sig at estimere den med:

$$\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$

hvor vi nu kun summerer over stikprøvens elementer. Dette estimat ville faktisk *i middel* give den rigtige varians σ^2, såfremt vi foretog en midling over alle tænkelige stikprøver. Problemet er bare, at vi ikke kender μ, så det bedste, vi kan gøre, er at benytte den empiriske middelværdi $\overline{x}$ som et estimat for μ. Derved har vi

$$s_n^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}$$

Desværre har den tilhørende estimator bias. Hvorfor dette er tilfældet, er vist smukt på en Wikipedia side med titlen: *Bessel's Correction*. Her vises det nemlig, at der gælder:

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 \le \sum_{i=1}^{n}(x_i-\mu)^2$$

og der gælder kun lighedstegn, såfremt $\overline{x}=\mu$. Denne ulighed kan fås frem ved at bruge noget så simpelt som kvadratet på en toleddet størrelse: $(a+b)^2 = a^2+2ab+b^2$. Man kan

se detaljerne på ovennævnte side. Det er ved at udskifte n med $n-1$, at vi har set, at den nye estimator bliver central. Herved får vi formel (2) for stikprøvevariansen:

$$s^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2 \quad \text{(Stikprøvevariansen)}$$

NB! Det skal siges, at vi i udgangspunktet ovenfor for simpelheds skyld – og for at notationen ikke bliver for tung – har antaget, at vi har et endeligt antal elementer N i populationen. Argumenterne kan også føres igennem, selv om populationen er repræsenteret ved en diskret fordeling med tælleligt mange elementer eller endda en kontinuert fordeling.

□

Pointe 4: Man kan undersøge rimeligheden af formel (2) for stikprøvevariansen ved at foretage *simuleringer* på en computer, dvs. programmere computeren til at udtrække et stort antal stikprøver på tilfældig vis, beregne stikprøvevariansen efter formlen (2), og så tage middelværdien af alle stikprøvevarianserne med henblik på at undersøge, om resultatet er tæt på σ^2. Alternativt kan man vælge at kigge på fine YouTube videoer, der beskriver sådanne simuleringer.

Pointe 5: Til slut skal det lige nævnes, at mens S^2 ifølge sætning 3 er en central estimator for variansen σ^2, så er kvadratroden af den faktisk *ikke* en central estimator for spredningen σ! Det er dog alligevel S man bruger som estimator for spredningen. Den leder til formlen (3) for stikprøvespredningen. Man kan læse om det på følgende Wikipedia side: *Unbiased estimation of standard deviation*.