Et debatoplæg: Om at formidle statistik

Birger Stjernholm Madsen,
bsm@novozymes.com

Mange matematiklærere underviser i statistik i gymnasiet. Det er min opfattelse, at formidling af statistik i gymnasiet dumper både på fremstillingen og på emnevalget. Dette debatoplæg er en sammenfatning af mine tanker, medens jeg skrev bogen Statistik for ikke-statistikere, Samfundslitteratur 2008.

Fremstillingen

Der bør efter min mening lægges vægt på en formulering, der er letforståelig frem for 100% matematisk stringens. Nogle få eksempler:

1. Om terminologi

Et af problemerne ved elementære publikationer om statistik er, at læseren præsenteres for adskillige ord for samme fænomen, ofte endda i samme publikation! Det gælder også helt elementære begreber.

Et eksempel: Det grundlæggende begreb, som illustreres af middelværdien, benævnes med adskillige forskellige betegnelser: (Mål for) niveau, position, beliggenhed, center, midtpunkt, lokalisering, tyngdepunkt, etc. Hvorfor har et så grundlæggende begreb så mange betegnelser?

Der er gjort forskellige forsøg på at standardisere statistiske betegnelser. Det mest vellykkede er nok ISO 3534 *Statistics – Vocabulary and symbols*, en standard i 3 dele. Der arbejdes på at oversætte denne standard til dansk.

Omvendt er det vel et rimeligt forlangende, at væsentlige begreber rent faktisk også har et navn! Lad os betragte formlen for et 95% konfidensinterval for en middelværdi i normalfordelingen:

$$\bar{x} \pm 2 \cdot \frac{s}{\sqrt{n}}$$

Denne formel gælder med god tilnærmelse, hvis antallet af observationer er rimeligt stort, f.eks. $n > 20$. Her er s standardafvigelsen eller "spredningen" i stikprøven.

Størrelsen $2 \cdot \frac{s}{\sqrt{n}}$ er af fundamental betydning for praktisk arbejde! En så fundamental størrelse har vel et navn? Ikke i internationale standarder. Heller ikke i de fleste lærebøger, hvor man kan finde betegnelser som "den halve længde af et 95% konfidensinterval" eller simpelthen "tallet efter ± ".

Inden for teorien for stikprøveundersøgelser kaldes denne størrelse i flere bøger simpelthen "den statistiske usikkerhed". At "ophøje" denne betegnelse til standardterminologi vil være et naturligt valg. Hvorfor er det så ikke sket?

2. Om græske bogstaver, symboler m.v.

Mange bøger om statistik starter med en tabel over det græske alfabet! Læseren efterlades med det indtryk, at dette virkelig må være væsentligt. Men: Hvor mange græske bogstaver har vi egentlig brug for?

Σ er som bekendt en sum, og det har de fleste heldigvis styr på takket være regnearkene!

μ og σ er jo henholdsvis middelværdi og spredning i "populationen", de er svære at undgå.

ISO 3534-2:2006

ISO 3534-2:2006 defines applied statistics terms, and expresses them in a conceptual framework in accordance with ISO normative terminology practice. Term entries are arranged thematically. An alphabetical index is provided. Standardized symbols and abbreviations are defined.

The two principal purposes of ISO 3534-2:2006 are, specifically, to establish a common vocabulary for use throughout ISO/TC 69 standards, together with the broader intent to enhance the preciseness, clarity and cohesiveness in the usage/application of applied statistics generally. The mathematical level has deliberately been kept to a low level in order for the content to be made readily comprehensible to the widest possible readership.

Fra www.iso.org/iso/catalogue_detail.htm?csnumber=40147. På denne hjemmeside kan denne ISO standard købes.

Hvad resten af de græske bogstaver angår: Vi har ikke brug for dem!

Også forskellige symboler vanskeliggør læsningen af statistiske publikationer. F.eks. vil vi professionelt ofte bruge en "hat" som symbol for estimat for parameteren p i en binomialfordeling og skrive

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

Måske burde vi i stedet blot skrive

$$p = \frac{x}{n}$$

Jeg ved godt, det ikke er korrekt! Men: gør det noget? Hvad med at opgive den matematiske præcision for at opnå større læselighed?

3. Om matematiske formler

Hvad skal vi med formler som f.eks.

$$f(x) = \frac{\exp\left(\frac{-x^2}{2}\right)}{\sqrt{2\pi}}$$

Denne formel for tætheden for normalfordelingen med middelværdi 0 og standardafvigelse 1 findes i adskillige "elementære" bøger om statistik! Hvorfor? Hvad skal vi med den? Hvilken informationsværdi indeholder den? Der er vel ingen, der udfører disse beregninger i hånden nu til dags? Og hvis man gør, så er ovenstående formel for resten ikke særlig velegnet, da det er fordelingsfunktionen, dvs. integralet, man har brug for til praktiske beregninger!

I øvrigt er det også bemærkelsesværdigt, så få anstrengelser, der gøres for at lette tilværelsen for læseren af de statistiske formler. Et eksempel:

I forbindelse med estimation af parameteren p i en binomialfordeling kan vi under forudsætning af simpel tilfældig udvælgelse, og at populationen er meget større end stikprøven, opstille følgende formel for den statistiske usikkerhed på estimatet $p = \frac{x}{n}$:

$$u = 1{,}96 \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

Under forudsætning af, at p ikke er alt for langt fra 0,5 f.eks. $0{,}3 < p < 0{,}7$ og at vi ignorerer den ubetydelige forskel mellem 1,96 og 2, reduceres dette til formlen:

$$u = \frac{1}{\sqrt{n}}$$

Denne formel må siges at være bemærkelsesværdig simpel! Og den er samtidig ekstremt nyttig! Alligevel er det ikke lykkedes mig at finde den i én eneste bog om statistik. Hvorfor?

Emnevalget

Emnevalget bør styres af, at hvilke emner, der er nyttige i det praktiske arbejde med statistik. Dels er der efter min mening et par spøgelser, der bør aflives, dels er der emner, der traditionelt lægges for lidt vægt på i elementære bøger om statistik. Nogle eksempler:

A. Sandsynlighedsregning

I mange bøger om statistik finder man en introduktion til sandsynlighedsregning. Det hører naturligvis med til statistikerens tankegods, men hvad skal vi med formler som

$$P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

Er der nogen, som har nytte af denne information? Hvad skal vi bruge den til? Bidrager formlen til at forstå binomialfordelingen?

Hvis man absolut skal introducere kombinatorik og sandsynlighedsregning, så gør det som en matematisk øvelse, ikke som en introduktion til statistik!

B. Konfidensintervaller versus test

Praktikeren, der skal sammenligne middelværdien i to grupper, kan vælge at udføre den klassiske t-test. Men mange har svært ved at forholde sig til test af hypoteser, p-værdier, signifikansniveau etc. Det er svært stof begrebsmæssigt!

Men han kan også beregne et konfidensinterval for forskellen mellem middelværdierne, det vil mange have langt lettere ved at forholde sig til.

Dette kan naturligvis også udnyttes i forbindelse med den praktiske planlægning af forsøg eller stikprøveundersøgelser. Praktikeren vil som regel godt kunne sige, hvor stor en statistisk usikkerhed, han kan leve med, f.eks. for forskellen mellem to middelværdier. Hvis han så også har et bud på, hvor stor den tilfældige variation er,

kan man finde ud af, hvor stort et forsøg, man skal lave.

C. Avancerede og utraditionelle emner

Man skal efter min mening ikke holde sig tilbage for utraditionelle emner. Et eksempel er en omtale af "skævhed" og "kurtosis", af og til kaldet "topstejlhed". Dette er traditionelt henvist til meget avancerede lærebøger, selv om de er nyttige redskaber for praktikeren. De kan hjælpe til afgøre, om data er normalfordelt. Endvidere er de indbygget i de fleste regneark.

Her bliver man naturligvis nødt til at have nogle vejledende grænser for, hvor store afvigelser fra 0, som maksimalt kan accepteres. Det er i bund og grund et meget simpelt spørgsmål! Men man skal lede meget længe for at finde et svar.

Hvis man leder længe i avancerede lærebøger, finder man nogle asymptotiske grænser. Fx er den maksimalt acceptable afvigelse fra 0 for skævheden givet ved udtrykket

$$2 \cdot \sqrt{\frac{6}{n}}$$

For kurtosis er grænsen det dobbelte. Men man finder ingen vejledning om, hvor stor stikprøvestørrelsen n skal være, før disse grænser kan bruges. Og så er man lige vidt.

Så må man ty til et af de værktøjer, som også kan formidles til ikke-fagfolk: Simulering!

Simuleringsstudier kan nemt foretages i et regneark. På denne måde kan man vise, at ovennævnte grænser gælder fint for små stikprøver, fx $n = 25$, for skævheden.

Derimod skal vi væsentligt højere op, før de fundne grænser gælder for kurtosis, helt op i nabolaget af $n = 1000$. For mindre værdier af n kan man ved hjælp af simulering beregne vejledende grænser for kurtosis. Dette er et godt eksempel på, at resultater af simuleringsstudier kan give nyttig information!

Kommentarer til ovenstående betragtninger er velkomne! ◊