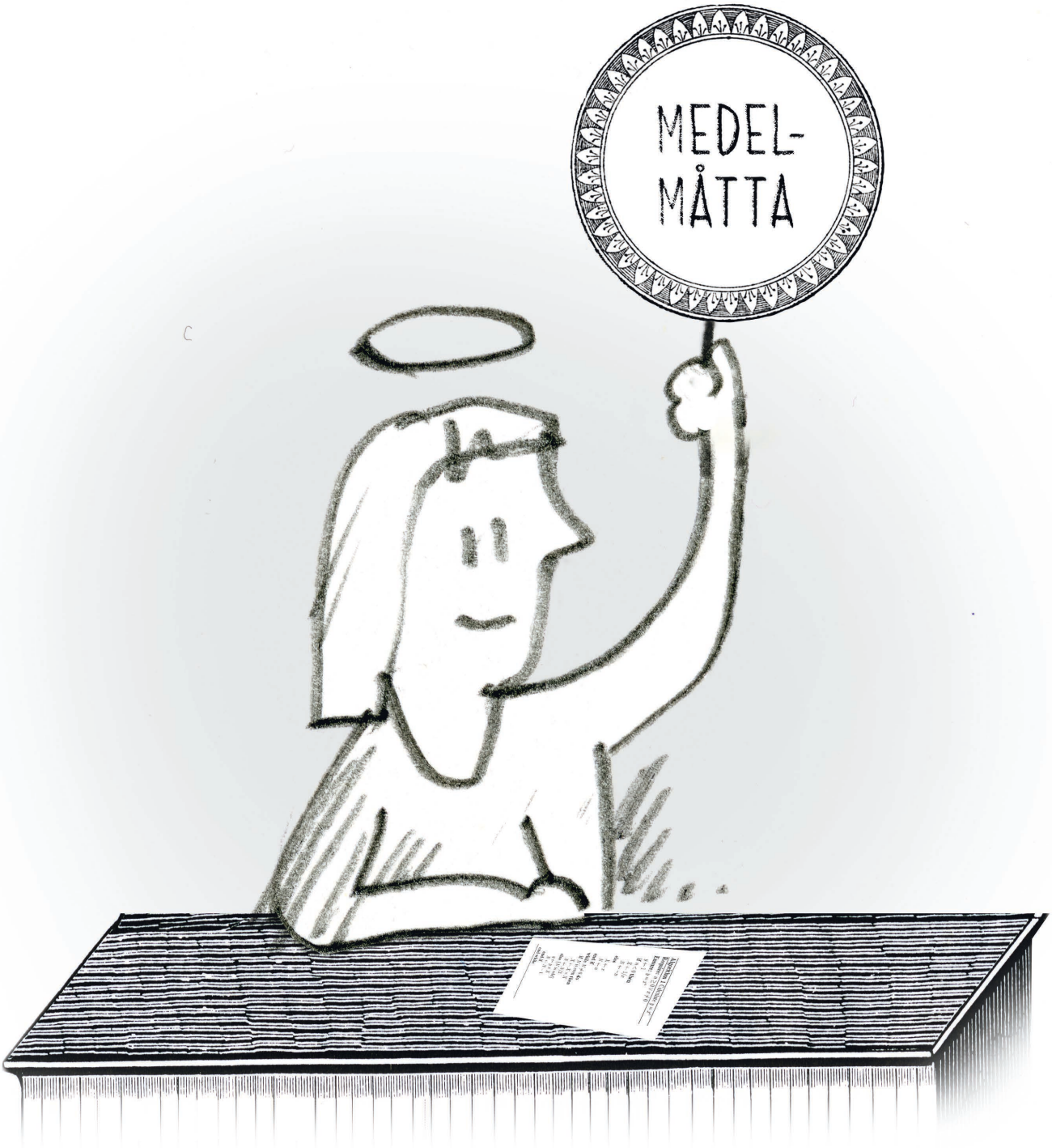


Medelvärde, median eller medelmåtta? Lägesmått är en djungel om man inte har järnkoll på vad som gäller! Illustration: Anders Gunér.

Statistikens atomer och molekyler

Statistikens minsta beståndsdelar är förstås data. Och det allra enklaste vi kan beräkna är medelvärde, median, typvärde och range. Okej, ni vet redan vad det är? Då kan ni skippa den här artikeln, men inte förrän ni tvärsäkert analyserar problemet med min dotters matteprov …

Vad tror de om oss föräldrar?
Min dotter kommer hem med resultatet från sitt första matteprov på gymnasiet. Glatt meddelar hon att hon hade ett resultat över medel; 19 poäng. Jag läser meddelandet från skolan: Max 24 poäng, medel 17 poäng och median 21 poäng. Och min dotter hade alltså 19 poäng, över medel men under median. Jag ser inte riktigt lika ljust på situationen som min dotter. Du som redan blixtsnabbt analyserat situationen (och funderat på lämpliga åtgärder) kan skippa resten av artikeln. Ni andra: Simma på.

Hur ser data ut?
Det absolut första man gör när man analyserar data är att SE PÅ DATA, ibland är det förstås inte möjligt, när ett dataset innehåller miljontals observationer (registerstudier) men ofta kan man visuellt inspektera data. Ofta gör man enkla grafer och tittar på till exempel fördelning och eventuella avvikande värden (outliers). Nästa steg är att se på medelvärde, median och kanske spridningsmått som varians eller minsta och största värde. Det låter ju väldigt trivialt, men många av de värsta felsteg jag sett handlar just om detta; att man inte tar hänsyn till hur data ser ut innan man påbörjar sina analyser. Så hur kan det bli så galet, bara för att man inte tar hänsyn till hur data ser ut?

Det låter ju väldigt trivialt, men många av de värsta felsteg jag sett handlar just om detta; att man inte tar hänsyn till hur data ser ut innan man påbörjar sina analyser.

Medelvärde versus median
Tänk att vi vill undersöka om en kräm förbättrar hudkvalitén jämfört med en placebokräm. Vi säger att de som får den aktiva behandlingen upplever en förbättring på 0, 0, 1, 3, 16 enheter och de som får placebo upplever en förbättring på 0, 0, 1, 2, 2 enheter. Ser man enbart på medelvärdet så är förbättringen för respektive grupp: 4 enheter (aktiv) och 1 enhet (placebo). Eftersom vi inspekterat data så förstår vi omedelbart att skillnaden i medelvärde beror enbart på en outlier; patienten som upplevde en förbättring på 16 enheter. Det var en stor skillnad i medelvärde: 4 enheter jämfört med 1, median är däremot densamma för bägge grupper, nämligen 1. En snabb titta på data gör att vi förstår att data inte är normalfördelade och att till exempel medelvärde inte är ett bra mått. Ett annat lägesmått som används ibland är typvärde, det vill säga det vanligaste värdet i ett dataset. I exemplet ovan är 0 typvärdet för både aktiv och placebo.

Så när skall man använda vad?
En första enkel ”regel” är att när medelvärde och median är lika stora så är data symmetriska. När data är symmetriska kan man ofta anta normalfördelning (även om det inte nödvändigtvis är sant) och resultaten av olika statistiska test, till exempel det allra mest använda testet, t-testet, fungerar bra. Om medelvärdet och median är olika så kan det betyda att vi måste använda andra statistiska metoder, som kan hantera att vi inte kan anta normalfördelning. Om data är asymmetriskt ”skeva”, som i exemplet ovan, kan det göra att man drar felaktiga slutsatser, till exempel att det verkar vara en skillnad i effekt mellan aktiv och placebo, men att det egentligen beror på några få outliers.

Ibland är median enda möjligheten!
Alla som är vana att läsa studier inom onkologi och hematologi vet att nästan utan undantag refereras det till

– data, lägesmått och variabilitet!

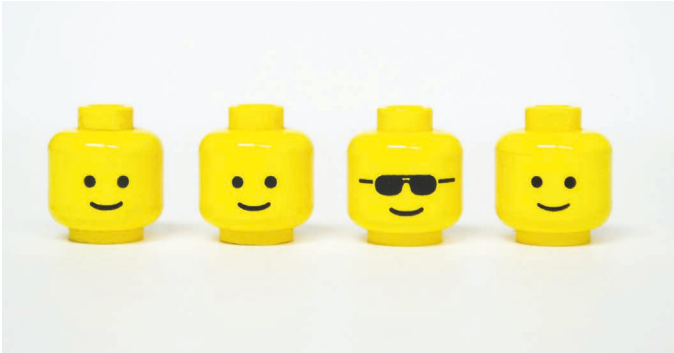
A ▶

Ellermore är Gullers

Att vi alla nu är en del av Gullers Grupp innebär idel fördelar, både för dig som kund och oss. Alla vi från Ellermore och Qre finns kvar, men nu finns det även möjlighet att utnyttja de nya fördelarna som en byrå med drygt 200 medarbetare och kontor i Stockholm, Göteborg, Malmö, Umeå och Sundsvall kan erbjuda. Till exempel får du allt från digital strategi, PR till gestaltning under samma tak. Parar du ihop det med Ellermore och Qres väldokumenterade styrkor inom läkemedelskommunikation, internationell B2B och konsumentreklam så kan vi få en riktigt spännande framtid tillsammans.

Vik sidan så att pilarna möter varandra.

medianöverlevnad och inte medelöverlevnad. Varför är det så? När man tänker på det så är det ganska logiskt. För att beräkna ett medelvärde behöver vi alla observationer, för att beräkna en median behöver vi hälften, de minsta (eller högsta värdena). Det betyder att vi kan beräkna medianöverlevnad även om den kliniska studien pågår och väldigt många patienter inte har fått återfall av sin cancer och fortfarande lever, vi behöver ju bara hälften (de kortaste) överlevnadstiderna. Sedan är det ock-



Standardavvikelse är bara en bra beskrivning på variabilitet när data är någorlunda normalfördelade.

så ofta så att en del patienter får återfall ganska tidigt, medan en andel patienter lever väldigt länge eller till och med kureras från sin cancersjukdom. Det betyder att överlevnadstiderna också blir skeva med en ”svans” mot höger, mot långa tider. I den här situationen passar alltså median bäst av flera orsaker.

Hur beskriver man variation?

Ofta vill man beskriva inte bara vad medianen eller medelvärdet är, utan också ge någon information om hur värdena sprider sig. Det mest använda måttet där är standardavvikelse som väldigt enkelt kan beskrivas som medelavståndet mellan de enskilda observationerna och medelvärdet. Det är egentligen lite mer invecklat (förstås) men som en generalisering duger det. Standardavvikelse är bara en bra beskrivning på variabilitet när data är någorlunda normalfördelade. Om data är ”onormala” kan det vara bättre att använda min och max, ofta kallas det ”range” men om man lusläser statistisk litteratur verkar det som om range egentligen är

B ▶

skillnaden mellan största och minsta värde – något som kallas variationsbredd på svenska. Ett annat sätt att beskriva variation i ett onormalt dataset är att dela in data i fjärdedelar. Första kvartilen är då värdet som avdelar de nedersta 25 procent av observationerna, andra kvartilen är naturligtvis detsamma som medianen och tredje kvartil delar av de översta 25 procent.

Allt detta tjat om normalfördelning…

Alla som träffat en statistiker vet att det ofta är en massa prat om huruvida data är normalfördelade eller inte. Varför är det så himla viktigt? Det korta svaret är att många statistiska analyser och test bygger på att man kan anta att data följer en fördelning som man kan beskriva matematiskt, till exempel normalfördelning. När det är möjligt får vi bättre statistiska test och mer statistisk ”power” i våra studier. När data är lite mer oförutsägbart fördelade, måste vi använda metoder som baserar sig på att vi endast använder ordningsföljden på observationer, att man använder ordningsföljd istället för de faktiska värdena. De analyserna blir också bra, men ofta får man lägre statistisk power och kanske då måste göra en större studie för att kompensera för att data inte kan antas ha en känd fördelning.

Hur var det då med matteprovet?

Eftersom medianen ligger nära maxpoängen betyder det att det var ett ganska lätt prov, hälften av alla elever hade en poäng på 21 poäng eller över (upp till 24 poäng). Att medelvärdet ligger lägre än medianen betyder att data har en ”svans” mot lägre värden som drar ner medelvärdet. Eftersom de flesta eleverna låg ganska nära maxpoängen så finns det naturligtvis ett större utrymme för att låga poäng kan dra ner medelvärdet en bit nedanför medianen. Vår analys av situationen är därför att några riktigt dåliga resultat dragit ner medelvärdet en bit under medianen. I det här fallet är det troligtvis lämplig att jämföra sin prestation mot medianen hellre än medelvärdet. Vi tackar för skolans förtroende för oss föräldrar; att vi har tillräcklig kunskap i statistik för att meningsfullt utvärdera våra barns resultat mot både medelvärden och medianer. Min dotter fick jobba lite mer med matteläxorna …

ANNA TÖRNER
Statistiker och verkställande direktör
Scandinavian Development Services