

Reduction of data (fig 1)

The use of average and standard deviation

The need to reduce the data

In practically all cases there is a need to reduce a set of data to some fewer numbers or other entities. The reduction is needed to grasp the main features of the data and its process without any significant loss of information. Some ways of reducing the data, each one with its good and bad features, are:

- a measure of location (most common - the average)
- a measure of variation (most common - the standard deviation)
- a graph (e.g. a histogram, a probability plot)
- a mathematical model (e.g. a probability model, a regression model)

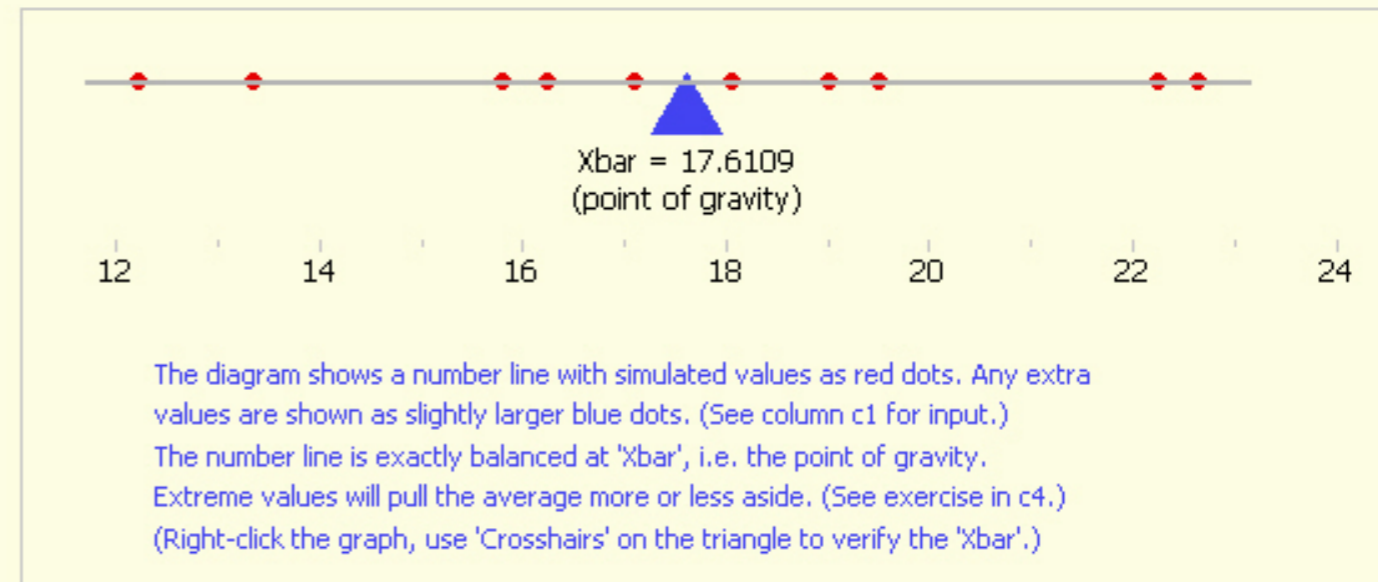
A measure of location - the average ('Xbar')

The average ("Xbar" - see formula to the right) is the most common way of reducing a set of data to one single number. (To be very clear: we talk about the arithmetic average as there are some other ones.)

The calculated Xbar-value has some different features and some of them are obvious and some less so. One bad feature might be that the calculated value seldom is a value amongst the measurements ('5.2 phone calls per day' is not a real value that has been recorded.)

Some features of the calculated average ('Xbar')

- The 'average' is calculated from the data.
- The 'average' is the 'point of gravity' of the data.
- The 'average' is often called 'Xbar'.
- The following formula is used: $\bar{X} = \text{sum}(X)/n$ ('n' is the sample size).
- $\text{Minimum} < \bar{X} < \text{maximum}$.
- The sort (mm, days, etc) of Xbar is always the same as the original data.
- The variation of Xbar, from sample to sample, is less than the variation of the original data.
- The 'average' estimates the true but unknown average of the process. This unknown value is designated 'mu'.



$$\bar{X} = \frac{\text{Value}_1 + \text{Value}_2 + \text{Value}_3 + \dots + \text{Value}_n}{\text{Number of values (n)}}$$

μ [mu:]>

'mu' is the theoretical mean and is only calculated using mathematical methods. 'mu' is never calculated from data. However, in order to get an estimate of 'mu', the 'Xbar' is used, thus 'mu' is estimated with a some uncertainty.

σ [sigma]>

'sigma' is the theoretical standard deviation and is only calculated using mathematical methods. 'sigma' is never calculated from data.

In order to get an estimate of 'sigma', the so-called sample standard deviation (s) is used. See fig 2.

Reduction of data (fig 2)

The use of average and standard deviation

The need to reduce the data

Apart from reducing all the data to one point (the average, fig 1) there is a need to describe the variation. In daily life we use practical, straight forward values as 'I need to wait 2 to 5 minutes before a bus comes' but when studying different kinds of processes we need a better measure of variation. 'Better' means here better mathematical features, better use of the existing data, etc. Unfortunately this also means a measure that takes some training to fully understand and use.

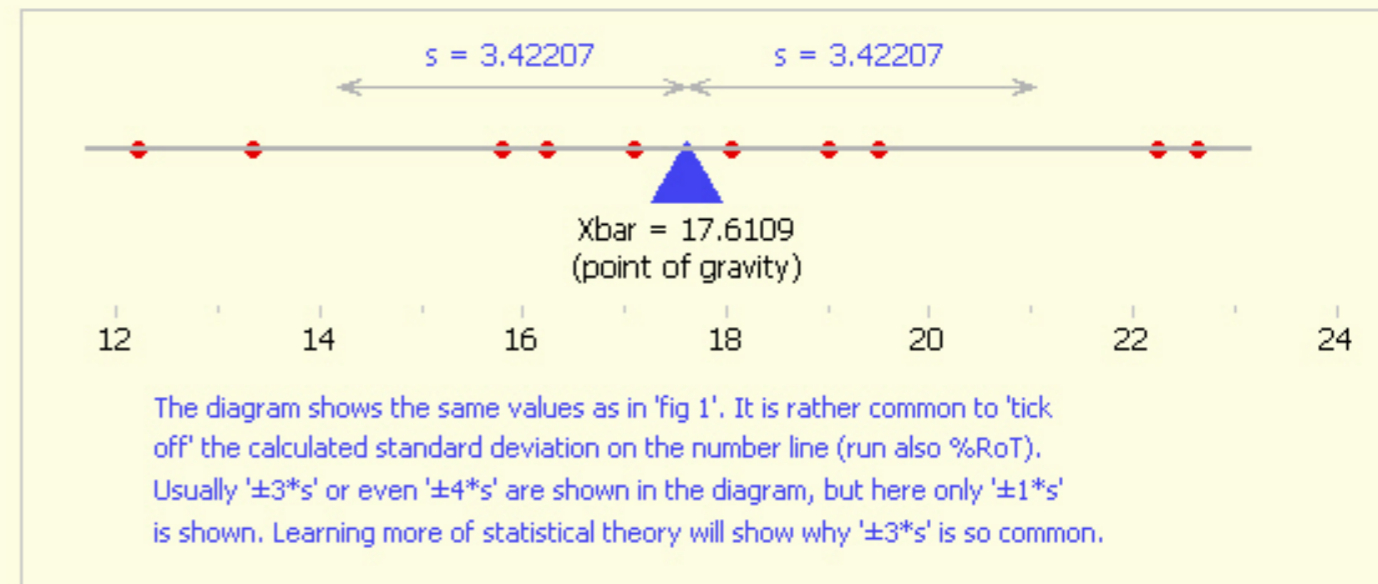
A measure of variation - the standard deviation ('s' or 'std')

The very calculation of the standard deviation is rather simple - 'calculate and square each distance from Xbar, sum all such squares, divide by (n-1) and take the square root of the result' (see e.g. ch 4 of 'A course in statistics'.) In all decisions based on data, the variation must be considered. Usually this will be done using applicable mathematics for the problem in question. However, the standard deviation can be used in several other ways, something that becomes more natural when learning more statistical theory. NB we distinguish between 's' (based on real data) and 'sigma' (based on pure mathematics only) and we say that 's' is an estimator of 'sigma'.

Some features of the calculated standard deviation ('s' or 'std')

- The 'std' is a measure of variation.
- The 'std' is calculated from the data.
- The 'std' is often called 's'.
- The following formula is used: $s = \sqrt{\frac{\sum(X - \bar{X})^2}{(n-1)}}$
- $0 < \text{'std'}$ (i.e. always positive).
- The sort (mm, days, etc) of 'std' is always the same as the original data.
- The 'std' estimates the true but unknown standard deviation of the process. This unknown value is designated 'sigma'.

(See www.ing-stat.se button [Articles] and the headline "What is sigma?" for an extensive discussion of the concept of 'sigma'.)



$$s = \sqrt{\frac{(X1-\bar{X})^{**2} + (X2-\bar{X})^{**2} + (X3-\bar{X})^{**2} + \dots}{(n - 1)}}$$

μ [mu:]>

'mu' is the theoretical mean and is only calculated using mathematical methods. 'mu' is never calculated from data. However, in order to get an estimate of 'mu', the 'Xbar' is used, thus 'mu' is estimated with some uncertainty (fig 1).

σ [sigma]>

'sigma' is the theoretical standard deviation and is only calculated using mathematical methods. 'sigma' is never calculated from data.

In order to get an estimate of 'sigma', the so-called sample standard deviation (s) is used.

Några exempel

Ett företag präglas av en mängd olika faktorer, motriktade viljor, skilda synsätt och annat som processer och människor i samverkan består av. I en stor mängd av dessa möten används ofta numeriska resultat för att förstå eller understryka vitala ståndpunkter eller argument.

Ibland 'låter man siffrorna tala för sig själv' vilket bara är sant i triviala eller självklara situationer. Statistiskt arbete och ett statistiskt angreppssätt präglas av en stark samverkan mellan metodkunskap i statistisk analys, ett synsätt på data och mätresultat samt hantering av dator och lämpligt datorprogram.

Dessutom krävs ett visst mått av nytänkande för att styra bort från slentrianmässig redovisning oavsett det gäller i sifferform eller med hjälp av grafer.

- **"Du påstår att vi genom att byta hårdare kan minska härdningstiden?"**

Låt oss genomföra ett försök som eliminerar osäkerheten. Låt oss göra detta med statistisk metodik och med de mätmetoder som vi själva använder. Genom att lägga upp det på ett smart sätt kan vi få med många parametrar och med få mätobjekt. Vi eliminerar också vårt eget tyckande och låter processen tala!"

- **"Vi har fått ett brev från Volvo som säger att det är för stor variation i ABC-produkten**

Volvo har skickat oss ett brev som säger att variationen i en av våra produkter, som vi levererar till dem, är för hög. De vill att vi presenterar våra egna siffror tillsammans med en grundläggande statistisk analys gjord med 7QCT ('seven Quality Control Tools'). De vill också att vi beskriver en åtgärdsplan och hur vi kommer att följa upp förbättringarna".

- **"Finns det egentligen något samband mellan X och Y-resultaten?"**

På våra möten sägs det ofta att det finns ett samband mellan antal partiklar i luften och antal felaktiga komponenter som produceras. Är det ingen som kan göra en trovärdig undersökning av detta?"

- **"Vad innebär ditt förslag att byta till elektronikkomponent A?"**

Vi har ju en relativt komplicerad produkt. Nu finns det ett förslag att om vi köper en dyrare komponent, men som har mindre toleransområde, så kommer vårt efterjusteringsarbete att minska. Jag vill se en ordentlig utredning som visar att totalvariationen och kostnaderna minskar!" (Eftersom komponenten skall ingå i en relativt komplicerad produkt är dator-simulering ett enkelt och billigt alternativ.)

Några exempel *(forts)*

- **”Vad innebär ditt förslag att byta till elektronikkomponent A?**
 Vi har ju en relativt komplicerad produkt. Nu finns det ett förslag att om vi köper en dyrare komponent, men som har mindre toleransområde, så kommer vårt efterjusteringsarbete att minska. Jag vill se en ordentlig utredning som visar att totalvariationen och kostnaderna minskar!” (Eftersom komponenten skall ingå i en relativt komplicerad produkt är dator-simulering ett enkelt och billigt alternativ.)

- **”Hur stort skall vårt reservdelslager vara?**
 Vi har idag levererat cirka 2700 enheter och vi vet att efterfrågan på nya tätningar är 1 per två år. Jag anser dock att vårt reservdelslager är för stort och därmed för dyrt. Å andra sidan vill jag inte att risken är att en kund står utan tätning då han behöver den. Jag vill dimensionera lagret så att risken att vi inte kan leverera en reservdel är, säg 1 på 1000. Är det någon som kan hjälpa mig?”

 - **”Vi har fått nya mål från huvudkontoret gällande tiden från orderstart till leverans**
 Igår fick vi nya mål för 2009. Som vanligt är de lite luddiga men jag tänkte att vi denna gång skall studera detta närmare. Det framgår inte vad menas med "max 10 dagar". Vi vet inte om det gäller 'i medeltal' eller 'alla', vi vet heller inte i detalj hur vi skall verifiera mål mot mätdata. Vi måste också bryta ned detta i delmål för att följa upp utfallet. Jag har pratat med lite folk som kan det där med statistik och nu har jag följande förslag: ...och på så sätt kan vi se hur vi framskrider...”.

- **”Vi har sedan i höstas mängder av data. Använder vi datamängden på rätt sätt?**
 Våra nyaste maskiner genererar en hel mängd data som vi helt enkelt inte använder. Ni bad om att få med detta i programvaran och då tycker jag att vi skall utnyttja detta. Jag är fullständigt övertygad om att många av våra problem kan lösas eller elimineras om vi bara brydde oss om vår egna data!” (Datadetektiven som genomför dataknådande. ”Datamining” är den engelska termen.)

Några exempel *(forts)*

- **”Varje vecka skickas det ut mängder med diagram, men ärligt talat så slänger jag dem**

Jag har hört att vi skall lägga in våra redovisningar på intranätet numera. Jag måste dock erkänna att jag inte tycker att våra grafer vare sig redovisar eller belyser våra problem. Ett antal procentstaplar, en för varje kvartal, gör ingen människa glad. Vi borde börja om och skapa något som är användbart och som även visar vårt största problem, variationen!”

- **”Jag vill att vår process för fel- och reklamationshantering förändras**

Vi får en hel del anmärkningar på reklamationshandlingen, det tar för lång tid. Jag är övertygad om att om vi utökar bemanningen så kan vi visa på besparingar i tid och böter för förseningar. Jag träffade igår en kille som kan det där med att simulera processer på dator. Han visade mig ett intressant exempel hur man kan pröva olika alternativ med utgångspunkt från våra egna data. Jag föreslår alltså att vi...”

- **Kundkrav, leverantörskrav, ISO-krav**

Det finns ofta en mängd kundkrav, leverantörskrav och dessutom ISO-krav på ett företag. För att kunna bemöta, föreslå eller uppfylla dessa krav måste man vara beredd på mäta, redovisa, argumentera eller visa att man klarar av en viss process eller åtagande.

Att förstå variation

Att kunna förstå och mäta, diskutera, hantera, bedöma, eliminera, etc, variation är oerhört viktigt då man vill förbättra och effektivisera processer. Till detta arbete finns det en mängd enkla verktyg.

Då en organisation börjar använda statistik, statistisk metodik och ett statistiskt synsätt brukar man se många positiva effekter:

- Ett gemensamt språk för att hantera data
- Fokusering på problemet och faktabaserade beslut i stället för tyckande eller pajkastning på andra avdelningar, chefer, konstruktörer, leverantörer, kunder, etc.
- Bättre samarbete mellan tillverkning och konstruktion
- Bättre integration av inköp och leverantörsfrågor och sammanhängande problem
- Bättre samarbete med underleverantörer och kunder
- Mindre slentrian när det gäller toleranser eller andra mål
- Bättre resonemang när det gäller mått och mätningar
- Applicering av metodik och synsätt på andra, mindre uppenbara, områden
- Bättre utnyttjande av personalens medverkan och kreativitet
- Bättre vaksamhet när det gäller löften, utsagor, m.m. när det gäller nya processer, nya metoder nya maskiner, etc
- Lättare att förklara varför man utnyttjar en kanske bättre metodik i stället för en annan som specificerats av kund i diverse kunddokument
- En beredskap att använda mer avancerad analysmetodik