F1F2F3 Introduktion. Statistisk undersökning. Deskriptiv statistik

Christian Tallberg

Avdelningen för Nationalekonomi/Statistik

Karlstads universitet

Vad är statistik?

1. Statistiska uppgifter. T ex som underlag och planering för beslut:
   - Kommunen som ska besluta om bostadsbyggandet måste ha tillgång till prognoser avseende folkmängdens framtida storlek och åldersfördelning
   - Företaget som planerar att ta fram en ny produkt måste ha en någorlunda realistisk uppskattning av marknaden för produkten.

2. Statistiska metoder att samla in, organisera och analysera data:
   - En statistisk utvärdering av ett nytt läkemedel
   - En marknadsundersökning
   - En befolkningsprognos

är exempel på ytterst skilda områden som kräver olika statistiska metoder.

Statistiska undersökningar

Klassificering efter mål eller syfte med undersökningen

1. Beskrivande undersökningar (deskriptiv statistik):
   - En folkräkning ger information om befolkningens storlek vid en viss tidpunkt.
   - Genom en statistisk kvalitetskontroll på en fabrik avslöjas hur stor andel av de tillverkade enheterna som inte uppfyller givna krav.

2. Analytiska undersökningar (någon form av (hypotes)test):
   - Har andelen moderater ökat under den senaste månaden?
   - Ger ett nytt läkemedel ett signifikant bättre resultat än ett visst äldre läkemedel?

Klassificering efter medel eller metoder vid datainsamling

1. Experimentella undersökningar: Ex. Vi vill studera sambandet mellan bromssträckans längd och hastigheten en bil håller.
   - Försöket kan upprepas önskat antal gånger. Detta innebär att vi kan göra så många observationer vi vill och därmed få så många mätvärden som behövs.
   - Försöksbetingelserna kan kontrolleras. Detta innebär att dessa kan hållas oförändrade (samma underlag i alla försök) eller ändras som vi själva önskar (bilens hastighet kan varieras).

2. Icke-experimentella undersökningar:
   - Konstatera vad som inträffat och i efterhand försöka kartlägga orsakerna. Ex. I vilken utsträckning kan löneskillnader på ett företag förklaras av faktorer som utbildning, anställningstid och kön?

	Me-	**-del**
	Experiment	Icke-experiment
Mål		
Beskrivning	Kvalitetskontroll	Folkräkning
Analys	Klinisk läkemedprö	Marknadsundersök

Typer av statistik:

- Deskriptiv statistik. Dvs metoder för att organisera, summera och presentera data på ett informativt sätt.
- Inferentiell statistik. Dvs metoder för att dra slutsatser om en population (populationsparametrar) med utgångspunkt från ett stickprov.

Genomförande av en statistisk undersökning

Innan vi genomför en statistisk undersökning måste följande tre övergripande frågor besvaras:

- *Vem* skall undersökas?
- *Hur* skall undersökningen göras?
- *Vad* skall undersökas?

Population och stickprov

I statistiken sägs ofta att data utgör ett *stickprov* från en *population*. Två typiska situationer när dessa två termer används är:

1. *Stickprov från ändlig population*.
   - Populationen är en tydligt definierad mängd av reellt existerande "element" (t ex människor, motorcyklar, aktieportföljer), som man vill ta reda på någonting om. Från en sådan population väljs genom något urvalsförfarande ett stickprov av "element". Data utgörs då av observerade värden på en eller flera undersökningsvariabler för de utvalda "elementen".

- *Exempel*: Vid en partisympatiundersökning. Ett stickprov på 1000 personer från en population av svenska medborgare 18 år eller äldre vid en viss tidpunkt.
- Sammanfattningsvis:

  Population = konkret mängd av element

  Stickprov = utvald delmängd av populationen

  Slumpen bestämmer vilket element som väljs.

2. *Stickprov från* (*en tänkt*) *oändlig population*.
   - Det finns en slumpmekanism eller slumpprocess (t ex en tillverkningsprocess) som genererar en följd av utfall på en slumpvariabel. Stickprovet utgörs här av den följd av värden som observeras. Sannolikhetsfördelningen för den observerade slumpvariabeln brukar i detta sammanhang kallas för en "oändlig population". Sannolikhetsfördelningen kan ju sägas beskriva fördelningen för de (oändligt många) möjliga observationer som skulle kunna genereras av den aktuella slumpprocessen (vi återkommer till detta längre fram i kursen).

- *Exempel*: 500 kast med en tärning ger värdena $x_1, x_2...$, där x_i är antalet prickar i det kastet. Kan beskrivas som att värdena är ett stickprov från en tänkt, oändligt stor population som utgörs av värdena 1, 2, ..., 6 som alla inträffar med samma sannolikhet. Dvs, den oändliga populationen är egentligen detsamma som sannolikhetsfördelningen för den slumpvariabel som observeras 500 gånger.

- *Exempel*: Med ett mätinstrument görs 50 upprepade mätningar av en och samma sträcka. De 50 mätningarna ger lite olika resultat från gång till gång pga ett mätfel, som antas uppträda slumpmässigt. Under antagandet att mätfelet är en normalfördelad variabel, så utgör de erhållna mätvärdena $x_1, x_2..., x_{50}$ ett stickprov från en tänkt oändligt stor population av normalfördelade "möjliga" mätvärden. Den oändliga populationen är alltså i detta fall en normalfördelning.

- Sammanfattningsvis:

  Population = sannolikhetsfördelning för den slumpvariabel som observeras

  Stickprov = den följd av värden som *observeras* på slumpvariabeln

  Slumpen bestämmer vilka variabelvärden som genereras.

- *Vem* skall undersökas?

Definition av en ram:

- Register eller annan förteckning över populationens element.

- *Hur* skall undersökningen göras?

Val mellan total - och urvalsundersökning.

Totalundersökning

Varför totalundersökning?

- Man är intresserad av att få en redovisning av resultaten i små delgrupper av populationen. Vid urval är risken stor att vissa (intressanta) grupper inte representeras av några personer.

- Populationen är så liten att det är omotiverat att göra annat än en totalundersökning.

Urvalsundersökning

Varför ett stickprov?

- Tidskrävande
- Kostsamt
- Destruktiva omständigheter kring vissa test
- Resultat från stickprov är adekvata

Vad kan vi påverka?

- Hur skattningen bestäms
- Hur stort urval som görs
- Hur observationerna samlas in (urvalsdesign).

Urvalsmetoder

Det finns olika urvalsmetoder (eller *urvalsdesign*) som brukar klassificeras i följande två grupper:

- *Icke-sannolikhetsurval* T. ex de första 300 individerna man möter på gatan. Ger dåliga skattningar om en hel stad eller ett helt lands invånare är populationen. Dessutom finns ingen (känd) slumpmekanism med i urvalsförfarandet.
- *Sannolikhetsurval* innebär att varje element i *populationen* har en *känd* sannolikhet som är större än noll att komma med i urvalet. Innebär att man får en samplingfördelning för skattningar, och får därmed en uppfattning om hur pass bra eller säkra skattningarna är (genom att göra t ex intervallskattningar) (vid icke-sannolikhetsurval kan man få en fördelning för skattningar också, men subjektiva modellantaganden krävs då).

Några vanliga sannolikhetsurvalsmetoder:

- *Obundet slumpmässigt urval* (OSU)

Ett OSU innebär att varje individ i populationen har samma sannolikhet att väljas till stickprovet.

Fördel: Relativt enkelt att genomföra.

Nackdelar:

- Man kan missa (en grupp av) individer som har avvikande variabelvärden. Det kan ge upphov till skeva skattningar.
- Man utnyttjar ingen relevant *apriori* information om individerna Det medför att precisionen i skattningarna kan bli sämre.

- *Stratifierat urval*

Populationen delas in i ett lämpligt antal grupper baserat på någon egenskap (stratifieringsvariabel) som korrelerar med undersökningsvariabeln. Sedan dras ett sannolikhetsurval av lämplig design från varje strata. Stratifieringen görs för att garantera att samtliga grupper i populationen representeras i stickprovet.

Fördel: Man får bättre precision i skattningen av populationsparametern (om det finns ett samband mellan stratifieringsvariabeln och undersökningsvariabeln).

Nackdel: (Kan vara) lite knöligare och mer tidskrävande att genomföra.

Exempel: Vi vill skatta medellönen i en population. Vi vet att kvinnor och män har olika lön för samma arbete. Dela in populationen i två grupper (strata), kvinnor och män. Dra ett sannolikhetsurval (t ex OSU) från varje strata. Lön är undersökningsvariabel och kön är stratifieringsvariabel.

- *Systematiskt urval*

Ett register över populationens individer delas in i n stycken intervall med a individer i varje (där $a = N/n$).

En första individ dras slumpmässigt, och med samma sannolikhet, från de a första individerna i registret.

Resten av stickprovet bestäms genom att systematiskt dra den $a : te$ individen.

Fördel: Enkelt att genomföra (endast en slumpmässig dragning).

Nackdel: Cyklisk trend som följer intervallängden ger upphov till skeva skattningar.

- *Gruppurval* (klusterurval)

  Populationens individer grupperas i delpopulationer (baserade på t ex geografisk spridning). Ett sannolikhetsurval (t ex OSU) dras av delpopulationerna. Ett stickprov av individer erhålls sedan genom att varje individ i urvalet av delpopulationerna undersöks, eller så dras ett sannolikhetsurval (t ex OSU) av individer från urvalet av delpopulationer.

  Exempel: Man vill göra en enkätundersökning om studenters alkohol och drogvanor. Först stratifieras skolorna efter geografiskt läge. Ett OSU av skolor dras från varje strata. Ett OSU av klasser dras från varje skola. Samtliga studenter i urvalet av klasser ingår i stickprovet och får svara på enkäten.

Fördelar:

- Ingen urvalsram (register) som identifierar varje individ krävs. Kan bli kostsamt att skapa en.

- Om individerna är spridda över ett stort geografiskt område kan det bli kostsamt att leta upp alla i stickprovet.

De förutsättningar som antas gälla fortsättningsvis på denna kurs är antingen

- att stickprovet har erhållits genom OSU från en *ändlig* population, som är *mycket stor* i förhållande till stickprovet

  eller

- att stickprovet utgörs av ett antal *oberoende observationer* från en tänkt *oändlig* population.

Vi kommer *inte* att behandla följande två fall (som i och för sig är viktiga):

- stickprovet har erhållits genom OSU från en *ändlig* population, som *inte* är mycket stor i förhållande till stickprovet

- stickprovet har erhållits genom någon *annan urvalsdesign* än OSU.

Olika insamlingsmetoder

- Primärdataundersökning: Att data samlas in för första gången.

- Sekundärdataundersökning: Man använder information och data som redan finns tillgängligt, och som kan ha samlats in för ett annat ändamål.

Olika insamlingsmetoder för primärdata:

- Enkäter
  1. Postenkäter
  2. Internetenkäter
  3. Gruppenkäter
  4. Besöksenkäter
- Inervjuer
  1. Besöksintervjuer
  2. Telefonintervjuer
- Bokföring och direkt observation

Konstruktion av frågeformulär

- *Vad* skall undersökas?

De frågeställningar som kan vara aktuella kan t ex vara:

- Hur många personer har en viss egenskap?
- Hur skiljer sig andelen med en viss egenskap i olika grupper?
- Hur har andelen med en viss egenskap förändrats sedan förra undersökningen?
- Hur ser sambandet mellan två variabler (eller egenskaper) ut?
- Hur ser framtiden ut med avseende på vissa variabler (eller egenskaper)? Dvs, vi vill göra prognoser.

Att tänka på vid formulering av frågor:

- Undvik ledande frågor utan gör dem så neutrala och balanserade som möjligt.
- Undvik värdeladdade eller prestigeladdade formuleringar.
- Undvik hypotetiska frågeställningar.
- Fråga om en sak i sänder.

- Använd ett korrekt, enkelt och lättfattligt språk.
- Läs frågan högt/Gör en provundersökning.
- Förklara förkortningar och fackuttryck - om de alls måste användas.
- Använd inte negeringar i frågan.
- Precisera frågan i tid och rum.
- Begränsa antalet frågor.

- Försök att, om möjligt, undvika öppna frågor. Knöligt att enkelt sammanställa resultatet.
- Försök att undvika känsliga frågor. Svaren ej tillförlitliga.
- Ha balanserade svarsalternativ.

Vad är din inställning till reklam i statliga televisionen?
T ex "positiv","ganska positiv","varken positiv eller negativ","ganska negativ","negativ", "vet ej".

- Ha alltid med ett svarslaternativ för osäkra respondenter. Detta kan exempelvis vara "vet ej", "har ej tagit ställning", "har ingen åsikt".

Kodning av svaren på attitydfrågor:

Vad är din inställning till reklam i statliga televisionen?
T ex "positiv" = 1,"ganska positiv" = 2,"varken positiv eller negativ" = 3,"ganska negativ" = 4,"negativ" = 5, "vet ej" = 5.

Kodning av värden på dikotoma (eller binära variabler eller 0/1 - variabler):

Kön: "Kvinna" = 1, "Man" = 0.

Olika feltyper i en undersökning

- *Urvalsfel*. Det fel som beror på att vi enbart gör ett urval som varierar från stickprov till stickprov. Det innebär att värden (på skattningar) i stickprovet oftast inte kommer att vara samma med värdena (på parametrar) i populationen.

- *Icke - urvalsfel*:

  1. *Täckningsfel*: Element ingår i ramen som inte ingår i populationen, eller tvärtom.
  2. *Mätfel*: Skillnaden mellan erhållet (uppgivet) och sant värde.
  3. *Bearbetningsfel*: Datorbearbetning, kodning.
  4. *Bortfallsfel*: Enheter i stickprovet som man ej fått något svar ifrån.

Deskriptiv statistik

Två centrala begrepp:

- *Data* = uppgifter (numeriska eller av annat slag) som har insamlats och registrerats, i syfte att bearbetas, analyseras och tolkas. Kan ligga till grund för slutsatser eller beslut.

- *Variabel* = en egenskap som kan variera mellan olika element i populationen. T ex kön, ålder, inkomst, antal barn.

Variabeltyper

Man brukar skilja mellan

- *kvantitativa (numeriskt mätbara) variabler* och

- *kvalitativa variabler (kategori* eller *icke-numeriska variabler).*

För kvantitativa variabler skiljer man också mellan

- *diskreta variabler* (variabeln kan bara anta vissa värden, oftast (men inte alltid) heltalsvärden) och

- *kontinuerliga variabler* (variabeln kan anta alla värden inom ett intervall, dvs teoretiskt sett *oändligt* många värden)

Exempel: En person är kvinna, 59 år, gift, 164 cm lång, väger 63 kg, har tre barn, är emot EMU och har betyget VG i en 10-poängs statistikkurs.

Variabel	**Variabelvärde**	**Variabeltyp**
kön	kvinna	kvalitativ
ålder	59	kvantitativ, kontinuerlig (ell
civilstånd	gift	kvalitativ
längd	164	kvantitativ, kontinuerlig
vikt	63	kvantitativ, kontinuerlig
antal barn	3	kvantitativ, diskret
inst. EMU	nej	kvalitativ
betyg	VG	kvalitativ

Skaltyp (eller datanivå)

Skaltypen bestäms av datanivån, och avgör vilka beräkning som är möjliga att genomföra. Följande fyra skaltyper (från lägsta till högsta datanivå) kan variabler delas in i:

1. *Nominalskala.* Kan enbart klassificera och räkna frekvenser för variabelvärdena. Går ej att rangordna variabelvärdena (kvalitativa variabler).

2. *Ordinalskala.* Det går att rangordna variabelvärdena. Det går dock ej att säga något om *skillnaden* mellan olika variabelvärden. Mätvärdens summa eller differens ger ingen meningsfull information.

3. *Intervallskala.* Vi kan ange skillnaden mellan variabelvärden. Summor och differenser ger meningsfull information. Exempel: Temperatur. Det är fem grader varmare idag än igår. Däremot är kvoten mellan mellan två variabelvärden meningslösa. Om det är tio grader idag och det var fem grader igår kan man inte säga att det är dubbelt så varmt idag. En *absolut nollpunkt* saknas.

4. *Kvotskala.* Variabelvärdena har en *absolut nollpunkt*. Exempel: Längd. Om jag är 180 cm och min morsa är 160 cm kan vi bilda kvoten $180/160 = 1.125$, dvs jag är 12.5 procent längre än min morsa.

Exempel:

Variabel	**Variabeltyp**	**Datanivå**
kön	kvalitativ	Nominal
ålder	kvantitativ, kontinuerlig (eller?)	Kvot
civilstånd	kvalitativ	Nominal
längd	kvantitativ, kontinuerlig	Kvot
vikt	kvantitativ, kontinuerlig	Kvot
antal barn	kvantitativ, diskret	Kvot
inst. EMU	kvalitativ	Nominal
betyg	kvalitativ	Ordinal

Tabeller och diagram

Frekvensfördelning:

- Ett sätt att klassificera data så att antalet individer (frekvenserna) anges i varje klass.

Kvalitativa variabler

Exempel: I ett fackförbund A finns 1787 medlemmar, 683 män och 1104 kvinnor. I en *frekvenstabell* redovisas fördelningen för variabeln kön så här:

Kön	**Antal**
Män	683
Kvinnor	1104
Totalt	1787

Exempel: Antag att ett fackförbund B har 6227 medlemmar, 1812 män och 4415 kvinnor. En tabell som visar två eller flera frekvensfördelningar samtidigt kallas *korstabell* och ser ut på följande sätt:

	Antal	**medlemmar**	
Kön	**A**	**B**	**Totalt**
Män	683	1812	2495
Kvinnor	1104	4415	5519
Totalt	1787	6227	8014

- Fördelningen för kvalitativa variabler redivisas lämplige med *stapeldiagram* eller *cirkeldiagram*.

Kvantitativa variabler som antar ett fåtal värden

Exempel: Frekvensfördelning över antal barn per familj.

Antal barn	**Antal familj**	**Andel familj**	**Andel familj**
x	**frekvens**	**rel. fre (%)**	**kumulat fre**
0	200	20	20
1	250	25	45
2	300	30	75
3	150	15	90
4	50	5	95
5	40	4	99
6	10	1	100
Totalt	1000	100	

- Fördelningen för kvantitativa variabler som antar ett fåtal värden redovisas lämpligen med *stolpdiagram*.

Hur gör man om variabeln kan anta många olika värden? Det kan bli väldigt många stolpar då.

Kvantitativa variabler som antar många olika värden

Klassindela observationerna så att observationer som är ungefär lika stora hamnar i samma klass.

Exempel: Fördelning av 497 män efter kroppslängd.

Längd (cm)	**Antal män**
155-159	1
160-164	19
165-169	32
170-174	96
175-179	122
180-184	113
185-189	71
190-194	40
195-199	13
200-204	3
Totalt	510

- Vi avrundar mätvärdena så att de män som är mellan 174.5 cm och 175.5 cm fått värdet 175 cm. Män som finns i klassen 175-179 är alltså mellan 174.5 cm och 179.5 cm långa. Värdet 174.5 är klassens *undre gräns* och 179.5 dess *övre gräns*. 179.5 är också den undre gränsen för nästa klass 180-184.

- Skillnaden mellan den övre och den undre gränsen för en klass kallas *klassbredden*. På denna kurs tar vi bara upp fallet där alla klasser har samma bredd, i detta fallet 5 cm.

- *Klassmitten* (som kan ses som ett approximativt värde för de individer som tillhör en viss klass) ges av undre gränsen plus halva klassbredden. I exemplet 157, 162, 167 osv.

- Hur *klassindelningen* ska göras och hur många klasser vi ska ha är en bedömningsfråga. 8-12 är ett vanligt antal. Klassindelningen gör materialet mer överskådligt men innebär också informationsförlust. *Ju färre klasser desto mer information förlorar vi*.

- Avrundningar kan ske på olika sätt. T ex ålder avrundas alltid nedåt till närmsta heltal. Dvs klassen 25-29 år innehåller de individer som fyllt 25 år men ännu inte 30 år. Klassgränserna i denna klass är därför 25 år och 30 år, klassbredden 5 år och klassmitten 27.5 år.

- Fördelningen för klassindelade observationer redovisas lämpligen med *histogram* eller *frekvenspolygon*.

Kumulativ frekvensfördelning

- De kumulativa frekvenserna erhålls genom att stegvis addera frekvenserna från det lägsta variabelvärdet till det högsta.

Exempel: Kumulativ frekvensfördelning över antal barn per hushåll.

Antal barn	**Antal hushåll**	**Andel hushåll**
x	**kumulativ fre**	**rel. kum.fre (%)**
0	200	20
1	450 = 200+250	45
2	750 = 450+300	75
3	900 = 750+150	90
4	950 = 900+50	95
5	990 = 950+40	99
6	1000 = 990+10	100

Kumulativa fördelningen för kvantitativa variabler som antar ett fåtal värden redovisas lämpligen med trappstegsdiagram.

Lägesmått (genomsnittsmått)

Ofta vill man åt två mått som beskriver väsentliga egenskaper i datamaterialet. Ett av dem är *lägesmåttet*.

- *Typvärdet*. Typvärdet motsvarar fördelningens maximum. Dvs, det *vanligaste* värdet, eller det värde som har den största frekvensen. Typvärdet kräver minst nominaldata.

Exempel: Vad är typvärdet av dessa nationaliteter? Fördelning över variabeln nationalitet.

Nationalitet	**Antal**
Dansk	12
Pakistanier	58
Kuban	19
Irakier	27

Variabelvärdet pakistanier är typvärdet.

- *Medianen*. Medianen motsvarar mittobservationens variabelvärde när data är *ordnat* i storleksordning. Medianen kräver minst ordinaldata.

  Exempel: Medianen vid intelligenstest. Om vi har följande fem observationer:

  102 128 98 103 107

Ordna data

98 102 103 107 128

Medianen är 103.

Om vi har följande sex observationer:

98 102 103 107 128 131

Medianen är medelvärdet av de två mittersta värdena, dvs

$$\frac{103+107}{2} = 105.$$

- *Percentilerna* delar observationerna i 100 lika stora delar. Vid stora datamaterial hittas därför medianen lätt genom ordningsnummret för den 50 : te percentilen, som ges av

$$L_P = (n+1)\frac{P}{100}.$$

- L_{25}, L_{50} och L_{75} delar materialet i fyra lika stora delar och kallas *kvartiler*. Medianen är alltså dessutom både 50:de percentilen och andra kvartilen.

- Medianen används ofta när data innehåller *extremvärden*, dvs kraftigt avvikande värden, som gör att andra genomsnittsmått , t ex medelvärdet, blir missvisande. Medianen är ett *robust* lägesmått. Man talar t ex ofta om medianlön i stället för medellön.

- *Aritmetiska medelvärdet*. Medelvärdet är summan av de observerade värdena dividerad med antalet observationer. Medelvärdet kräver minst intervalldata.

För att kunna beräkna medelvärdet krävs först lite notation. En vanlig beteckning på en variabel är X (stor bokstav). T ex variabeln längd kan vi beteckna med X. De värden som *observerats* i en *population* av storlek N betecknas däremot $x_1, x_2, ..., x_N$ (små bokstäver), där x_1 är längden hos första individen i populationen, x_2 längden hos andra individen i populationen osv.

Matematiskt kan man då uttrycka populationsmedelvärdet som

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},$$

där x_i är den $i:te$ individens variabelvärde, och μ är *populationsmedelvärdet*.

Om vi drar ett *stickprov* från populationen av storlek n, betecknar vi de observerade värdena i stickprovet med $x_1, x_2, ..., x_n$. Stickprovsmedelvärdet beräknas då som

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Exempel: I en population av $N =$ sex familjer finns $x_1 = 1, x_2 = 2, x_3 = 2, x_4 = 4, x_5 = 3, x_6 = 3$ barn.

Populationsmedelvärdet blir

$$\mu = \frac{1 + 2 + ... + 3}{6} = 2.5,$$

dvs i genomsnitt 2.5 barn per familj.

Vi drar ett stickprov av $n =$ tre familjer och får observationerna $x_1 = 2, x_2 = 2, x_3 = 3$ barn.

Stickprovsmedelvärdet blir då

$$\bar{x} = \frac{2 + 2 + 3}{3} = 2.33,$$

dvs i genomsnitt 2.33 barn per familj.

- Egenskaper i populationen kallas *parametrar* (t ex populationsmedelvärdet) och betecknas ofta med grekiska bokstäver.

- Egenskaper i stickprovet kallas *statistikor* (eller estimatorer eller skattningar när vi kommer till statistisk inferens)(t ex stickprovsmedelvärdet) och betecknas ofta med latinska bokstäver.

- Viktig egenskap för medelvärdet

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ \sum_{i=1}^{n} x_i - n\bar{x} &= 0 \\ \sum_{i=1}^{n} (x_i - \bar{x}) &= 0, \end{aligned}$$

dvs summan av avikelserna från varje värde till medelvärdet blir alltid noll.

- *Viktat medelvärde.* Om k är antalet värden som variabeln X kan anta, är ett vanligt sätt att beräkna ett viktat medelvärde är

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + ... + w_k x_k}{w},$$

där

$$w = w_1 + w_2 + ... + w_k,$$

vilket innebär att summan av vikterna w_i/w är

$$\frac{w_1 + w_2 + ... + w_k}{w} = 1.$$

Vikterna kan då tolkas som relativa frekvenser.

Exempel: Vi drar ett stickprov på 1000 familjer.

Antal barn	**Antal familj**	**Andel familj**
x	**frekvens**	**rel. fre (%)**
0	200	20
1	250	25
2	300	30
3	150	15
4	50	5
5	40	4
6	10	1
Totalt	1000	100

Om $w_1/w = 200/1000 = 0.2, w_2/w = 250/1000 = 0.25$ osv blir genomsnittliga antalet barn per familj då

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{6} w_i x_i}{w} \\ &= 0.2 \cdot 0 + 0.25 \cdot 1 + ... + 0.01 \cdot 6 \\ &= 1.76 \end{aligned}$$

Spridningsmått

Det andra måttet som beskriver väsentliga egenskaper i datamaterialet är *spridningsmåttet.*

- *Varians* och *standardavvikelse.* Minst intervalldata krävs. Populationsvariansen ges av

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

dvs medelvärdet av de kvadrerade avvikelserna från populationsmedelvärdet. Populationsstandardavvikelsen ges av

$$\sigma = sqrt\left(\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}\right),$$

dvs roten ur populationsvariansen.

Intressantare och ett *nödvändigt* mått vid statistiska analyser (med hjälp av slumpmässiga urval) är stickprovets standardavvikelse som ges av

$$s = sqrt\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right),$$

dvs roten ur medelvärdet av de kvadrerade avvikelserna från *stickprovsmedelvärdet*.

Exempel: OBS!!! Jobba på detta exempel. Avkastningen (i tusen kronor) för en aktieportfölj vid sex slumpmässigt valda tidpunkter.

Resultat:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1.5	0.92	0.85
1	0.42	4
-0.5	-1.08	1
-2	-2.59	1
0.5	-0.08	4
3	2.41	25
$\sum x_i = 3.5$	$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 44$

Notera att *summan av avvikelserna* till medelvärdet är noll.

Standardavvikelsen blir då

$$s = sqrt\left(\frac{44}{6-1}\right) = sqrt\,(8.8) = 2.97.$$

Notera också att man dividerar kvadratsumman med $n-1$ (och inte med n). Detta har att göra med att för att kunna beräkna standardavvikelsen i stickprovet måste vi *först* beräkna medelvärdet *från stickprovet.* Vi "fuskar" då och straffas på detta sätt.

Oftast, särskilt om antalet observationer är många, är det enklast att använda följande *beräkningsformel*

$$s = sqrt\left(\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}\right).$$

Forts. exemplet:

$$\sum x_i^2 = 1^2 + 2^2 + 3^2 + 3^2 + 6^2 + 9^2 = 140$$

Standardavvikelsen blir

$$s = sqrt\left(\frac{140 - \frac{24^2}{6}}{6-1}\right) = sqrt\,(8.8) = 2.97.$$

Stickprovsvariansen, s^2, ges av standardavvikelsen i kvadrat.

Explorativ Data Analys (EDA)

Ytterligare sätt att åskådliggöra data.

- *Lådagram (boxplot)*. En (enligt min mening) alltför underskattad diagramtyp som på ett informativt sätt åskådliggör data. Visar första kvartilen, andra kvartilen (dvs medianen), tredje kvartilen, minsta värdet, största värdet, samt eventuella *extremvärden*.

- *Spridningsdiagram*. Antag att vi nu har *två* variabler. Dvs för varje individ har vi två mätvärden, t ex kön och inkomst. Vårt statistiska material består alltså av *observationspar*. Tidigare har vi beskrivit var och en av variablerna med hjälp av tabeller och diagram (vilket i och för sig är intressant). Det är också intressant att fråga
  - Finns det något samband mellan variablerna?
  - Hur ser i så fall detta samband ut?

Sambandet kan åskådliggöras med hjälp av ett spridningsdiagram.

Exempel: För åtta barn finns följande uppgifter om ålder och längd:

Barn	A	B	C	D	E	F	G	H
Ålder	1	2	3	4	5	6	7	8
Längd	68	91	102	107	105	114	115	127

Talparet markeras i ett koordinatsystem där man har den *oberoende* variabeln på x-axeln, och den *beroende* variabeln på y-axeln. Om vi har ett samband som inte enbart är numeriskt, utan har en verklighetsförankring (så kallad kausalitet), brukar värdet på den ena variabeln (den oberoende) styra värdet på den andra variabeln (den beroende). Dvs tolkning blir ju äldre man är desto längre är man (åtminstone upp till arton år), och inte tvärtom, ju längre man är desto äldre är man.