



Statistikens grunder 2 dagtid

HT 2012

F1 Deskription

Att beskriva, illustrera och sammanfatta en uppsättning observationer

Men först något kort om

- *Databildning (Nyquist kap 9)*

Databildning Kap 9

Data:

- Mätningar, observationer
- ex. 22 M 45 62 84

Metadata:

- Information om data, förklarar vad raden ovan står för, vad värdena betyder, utfallsrum mm.

Paradata (SCB och andra):

- uppgifter om datainsamlingen:
- Ex. 4 T 62 011

"Metaparadata":

- berättar vad paradata står för

Mätning 1

Validitet:

Mätprocessens förmåga att representera den studerade egenskapen.

Prediktiv validitet = förmåga att förutsäga observationer

Reliabilitet:

Mätprocessens förmåga att ge ungefär samma resultat vid upprepade mätningar.

Tillförlitlighet, slumpfel.

Mätning 2

Bias:

Ett **systematiskt** fel som introduceras av mätprocessen.

Slumpfel:

Ett **stokastiskt** fel som introduceras av mätprocessen.

$$Y = \mu + b + \varepsilon$$

Observerat värde

Sant värde

Bias

Slumpfel

Empiriska studier 1

Syfte: att på något sätt avgöra om en modell är en rimlig beskrivning av den verklighet vi studerar.

Experimentella studier:

- Objekt utsätts för olika behandlingar för studera effekter (respons)
- Randomiserad (slumpmässig) allokering av objekt till olika behandlingar
- Kontrollgrupper (placebo)
- Ofta för att studera orsakssamband (kausalitet)
- Etik?

Empiriska studier 2

Observationsstudier:

- Objekt väljs ut ur en population och studeras (mäts)
- Dessa får representera hela populationen
- Studera egenskaper i populationen och/eller jämförelser mellan olika grupper inom populationen
- Utsätts inte för en behandling, passivt observerande
- Svårt (omöjligt?) att dra absoluta slutsatser om kausalitet
- Etik?

Empiriska studier 3

Urvalsundersökningar:

- Kostnadsskäl – billigare undersöka ett fåtal jmf rt med alla
- Tidsskäl – går fortare (kostnad och aktualitet)
- Mer pengar över $\Rightarrow$ mätprocessen kan förfinas, bättre resultat jmf rt med totalundersökning
- Praktiskt omöjligt – t.ex. oändliga populationer
- Förstörande mätningar

Empiriska studier 4

Slumpmässiga urval:

- **Representativa** urval – vad menas med det?
- Aktivt välja urvalsobjekt som man anser vara representativa medför problem (kap 9 sid 21) – varför?
- Vad man "anser" behöver inte vara hela sanningen – inte ens delvis
- **Slumpmässiga** urval garanterar (om de görs rätt) vad man kallar **väntevärdesriktiga** skattningar med bra **precision**.

Deskription Kap 11-12

Att presentera och sammanfatta empiriska observationer (empirisk fördelning). **Explorativ analys!**

Att jämföra med en teoretisk modell (sannolikhetsfördelning).

- **Tabeller**
- **Grafisk framställning**

Men även

- **Sammanfattande mått**
 - medelvärde, median, andelar, standardavvikelse i en uppsättning empiriska observationer, data

Variabler

Repetition

- Kvantitativa variabler
 - Antar numeriska värden
- Kvalitativa variabler
 - Antar icke-numeriska värden
- Kontinuerliga variabler
 - Kan anta **samtliga** värden inom ett intervall
 - Kan vara ändlig eller oändlig
- Diskreta variabler
 - Kan anta **endast vissa** värden

Skalor

Repetition

Värdena som en variabel kan anta anges på olika skaltyper:

- Nominalskala
 - icke-numeriskt, latin nomen = namn
 - Ex. bilmärke, yrke m.m.
- Ordinalskala
 - Icke-numeriskt men kan ordnas
 - Ex. "bra, bättre, bäst"
- Intervallskala
 - Numeriska värden där avstånden är väldefinierade men inte kvoter
 - Ex. Celsiusskala
- Kvotskala
 - "20 är två ggr större än 10"

Repetition

Sammanfattning

Skaltyp	Variabeltyp		
	Diskret	Kontinuerlig	
Nominal	X	-	Kvalitativ
Ordinal	X	-	
Intervall	X	X	Kvantitativ
Kvot	X	X	

Olika variabeltyper och skalor kräver olika tabelltyper och diagramtyper!

Frekvenstabeller 1

- Kvalitativa – nominal och ordinal
- Kvantitativa – diskret och klassindelad kontinuerlig

Räkna antalen som faller inom varje definierad kategori

Nominal	Glassmak	Frekvens	Rel. frekv.	<i>Störst först!</i>
	Choklad	70	35,0 %	
	Vanilj	50	25,0 %	
	Jordgubb	45	22,5 %	
	Hallon	30	15,0 %	<i>Minst sist!</i>
	Lakrits	5	2,5 %	
	Summa	200	100 %	

Frekvenstabeller 2

Ordinal

Betyg	Frekvens	Rel. frekv.
Sämst	30	15,0 %
Dålig	55	22,5 %
OK	80	40,0 %
God	20	10,0 %
Bäst	15	7,5 %
Summa	200	100 %

Ordna efter rangordningen på skalan!

Diskret

Antal fel	Frekvens	Rel. frekv.
1	5	2,5 %
2	30	15,0 %
3	70	35,0 %
4	45	22,5 %
5	50	25,0 %
Summa	200	100 %

Ordna efter storleksordning!

Klassindelning

- När en variabel är kontinuerlig eller nästan kontinuerlig (hur många decimalers noggrannhet?)
- Gruppera närliggande till samma klass.
- **Klassbredd** måste definieras
 - Ex. 0-4,99; 5,00-9,99; 10,00-14,99
- Samma klassbredd eller olika?
 - Ex. åldersgrupper kan variera ibland
- Frekvenserna kan sammanställas klassvis i en tabell.

Grafisk framställning

Frekvenser (absoluta el. relativa)

- Stapeldiagram
 - Kvalitativa, nominal och ordinal
 - Ordna på motsvarande sätt som med frekvenstabeller
 - Uppdelade staplar
- Cirkeldiagram
 - Kvalitativa, nominal
- Stolpdigram
 - Kvantitativa, diskret
- Histogram
 - Kvantitativa, klassindelat kontinuerlig

Histogram

- Histogram används när man har kontinuerlig variabel.
- Samma som med tabeller, klassbredderna måste definieras.
- Frekvensen i en klass ska avspeglas i stapelns **area**, inte dess höjd!
- Om lika klassbredd kommer höjden bli proportionell mot frekvensen.
- Öppna klasser (ex. >65) markeras med streckade linjer
 - Eftersom vi inte vet var det slutar kan vi inte veta vad arean är!

Lägesmått 1

Empiriska lägesmått

- **Aritmetiskt medelvärde**

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

- Känsligt för extrema värden
- Ex. 2, 3, 4, 5 $\Rightarrow \bar{x} = 3,5$
- Ex. 2, 3, 4, 25 $\Rightarrow \bar{x} = 8,5$
- Stickprovsmedelvärde motsvaras av det teoretiska och abstrakta väntevärdet

Lägesmått 2

- **Medianen** delar ett data material i mitten, dvs. 50 % av observationerna ligger på vara sida om medianen
- Rangordna observationerna
 - n udda $\Rightarrow$ median = mittersta
 - n jämn $\Rightarrow$ median = medelvärdet av de två närmast mitten
- Ex. 2, 3, 4, 5 $\Rightarrow$ median = 3,5
- Ex. 2, 3, 4, 25 $\Rightarrow$ median = 3,5

Spridningsmått 1

Stickprovsvariansen

- (Nästan) samma definition som den teoretiska fördelningens varians
- Genomsnittligt kvadrerat avstånd till medelvärdet

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

*Obs! $n-1$ istf n
Medför bättre egenskaper*

Tchebysheffs olikhet

- k är ett tal s.a. $k \geq 1$
- För alla empiriska fördelningar gäller att andelen observationer som ligger i intervallet

$$(\bar{x} - ks, \bar{x} + ks)$$

är minst $1 - 1/k^2$

Ex.

$$k = 1 \Rightarrow \text{andelen} > 1 - 1/1^2 = 0$$

$$k = 2 \Rightarrow \text{andelen} > 1 - 1/2^2 = 0,75$$

$$k = 3 \Rightarrow \text{andelen} > 1 - 1/3^2 = 0,8889$$

$$k = 4 \Rightarrow \text{andelen} > 1 - 1/4^2 = 0,9375$$

Spridningsmått 2

- **Kvartiler** och **kvartilavstånd**
- $q_1 = 1$:a kvartilen
25 % nedanför, 75 % ovanför
- $q_3 = 3$:e kvartilen
75 % nedanför, 25 % ovanför
- Beräknas enligt samma mönster som medianen (se kap 11 sid 17)
- $IRQ = \text{Kvartilavstånd} = q_3 - q_1$
(eng. *interquartile range*)

Boxplottar 1

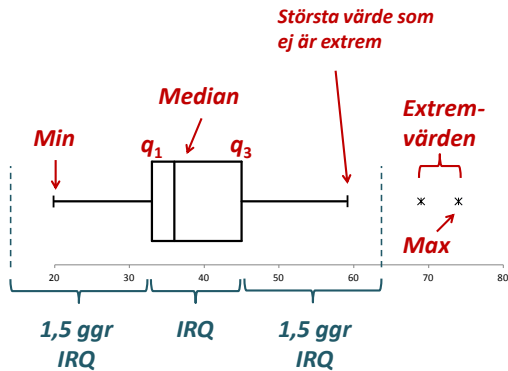
- Behöver
 - minsta och största värde
 - median, första och tredje kvartiler
 - definiera ev. **extremvärden**

Outliers: enligt en definition (Tukey) värden som ligger mer än 1,5 ggr IRQ till vänster om q_1 eller till höger om q_3 .

Extrema outliers: om avståndet är större 3 ggr IRQ

Boxplottar 2

- Exempel



Flera variabler

- Varför titta på flera samtidigt?
- Sambandsanalys
- Finns det stöd eller inte i data för ett samband (beroende) mellan olika variabler?
- Studera korstabeller och olika grafer/diagram

Frekvenstabeller 1

- Korstabeller**, man korsar två eller fler variabler.
- Absoluta eller relativa frekvenser
- Men ... nu kan vi välja:
 - **simultana** relativa frekv.
 - **betingande** relativa frekv.
- Med den senare är det oftare lättare att upptäcka samband.
- Se ex. kap 12 sid 5.

Beskrivande mått

- Vi studerade betingade fördelningar och dessas väntevärden och varianser.
- Det kan vi givetvis göra med empiriska data.
- Jämföra två eller flera grupper med **läge** och **spridning**.

Kvottabeller 1

- Besläktat begrepp (SCB): **magnitudtabeller**
- Isf medelvärden pratar man ofta (på SCB) om **totaler**
- Dvs. det uppsummerade eller **aggregerade** värdet för en variabel y betingat på värdet på en annan x
- Man kan förstås korsa flera förklaringsvariabler och aggregera en responsvariabel
 - Se t.ex. tabell 12.6 sid 10

Kvottabeller 2

För en redovisningsgrupp g :

- Kvot: medelvärde $\bar{x}_g = \frac{1}{n_g} \sum_{x_i \in g} x_i$
- Total: $t_g = n_g \bar{x}_g = \sum_{x_i \in g} x_i$
- Andel: specialfall av medelvärde.

$$p_{Ag} = \frac{1}{n_g} \sum_{x_i \in g} x_i$$

$$\text{där } x_i = \begin{cases} 1, & \text{om i harenegenskap} \\ 0, & \text{annars} \end{cases}$$

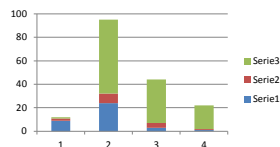
F2 Deskription forts

Grafisk framställning flera variabler

- Staplade staplar (kategoriska data)

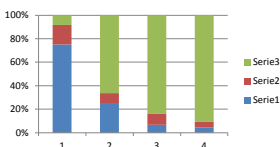
Absoluta

frekvenser
per grupp
(simultan fördelning)



Relativa

frekvenser
per grupp
(betingad fördelning)

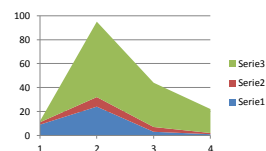


Staplade ytor

Variant av stapeldiagram

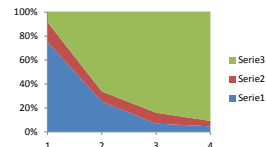
Absoluta

frekvenser
per grupp
(simultan fördelning)



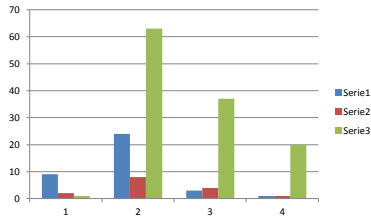
Relativa

frekvenser
per grupp
(betingad fördelning)



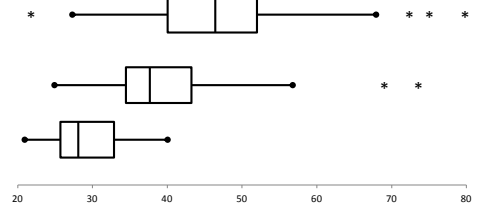
Stapeldiagram

- Grupperade staplar



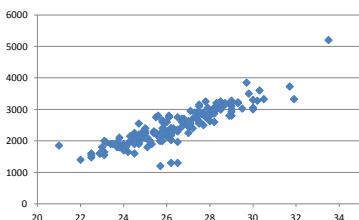
Boxplottar igen

Jämföra grupper



Punktplottar 1

- Varje talpar representeras av en punkt:



- Oerhört viktigt instrument när man studerar samband mellan (kontinuerliga) variabler!

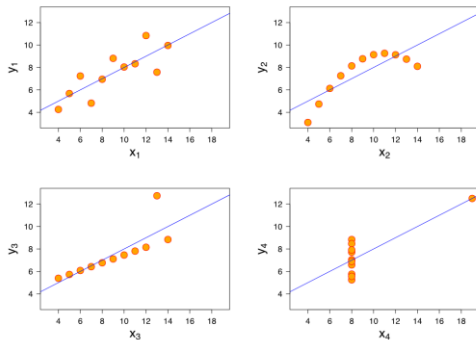
Punktplottar 2

Vad letar man efter?

- Starka eller svaga samband
- Positiva eller negativa samband
- Linjära eller icke-linjära samband
- Avvikande och extrema värden
 - huvudsakligen två typer men mer om det på nästa kurs!

Punktplottar 3

Ex. Anscomb's data set



Kovarians och korrelation

- Stickprovsmotsvarigheterna till kovarians och korrelation i en simultan bivariat fördelning

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

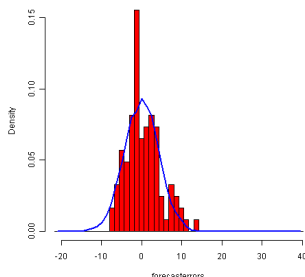
$$= \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right)$$

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Att jämföra med punktplottarna!

Jämföra empiri och modell

- Jämföra en empirisk fördelning med en teoretisk fördelning



- Den empiriska visas i ett histogram, den teoretiska som en graf

Fortsatt läsning

- Avsnitt 12.5 Samspelseffekt
- Avsnitt 12.6 Standardpopulationsmetoden
- Läs själva men särskilt 12.5 finns anledning att återkomma till när ni läser Regressionsanalys

Tidsserier

- När vi studerar en variabel över tid
- Betecknas ofta X_t
- Stort X för att är stokastiskt
- Index t för tid

Två sätt beskrivs i Nyquist

- Dekomponering, uppdelning i komponenter
- Stokastisk process

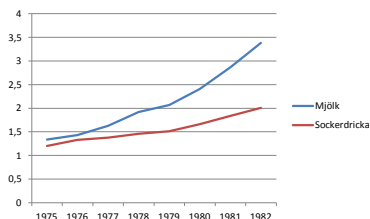
Tidsserier

- Vad utmärker tidsserieanalys jmfrt med tvärsnittsdata?
- Upprepade mätningar över tid
- **Beroende** observationer
- Frågeställningarna knutna till seriens **utveckling över tid**, t.e.x förändringar, trender och säsongsmässiga variationer mm.

(Se Kap 13 sid 4)

Grafisk framställning

- **Kurvdiagram**, tiden på x-axeln och observationerna mot y-axeln. Punkterna förbinds med linjer.



År	Mjolk	Sockerdricka
1975	1,34	1,2
1976	1,43	1,33
1977	1,63	1,38
1978	1,92	1,46
1979	2,07	1,51
1980	2,41	1,66
1981	2,87	1,84
1982	3,38	2,01

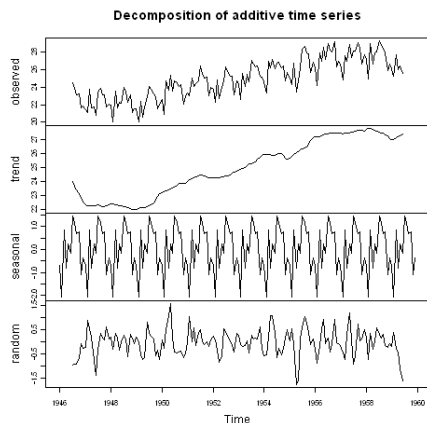
Komponenter 1

- Man tänker sig att varje observation består av fyra delar
 - Trend
 - Konjunktur
 - Säsong
 - Slump

$$X_t = Tr_t + K_t + S_t + \varepsilon_t$$

- Hur stor del var och bidrar med varierar
- Dels komponenternas utveckling över t och storlek

Dekomponering



- Multiplikativ modell:

$$X_t = T_t \cdot C_t \cdot S_t \cdot \varepsilon_t$$

Stokastiska processer 1

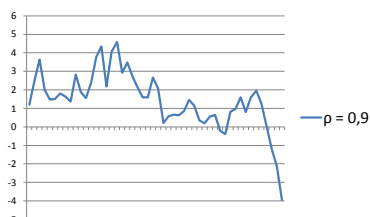
- Enklaste varianten:
1:a ordningens autoregressiv process, eller AR(1)

$$X_{t+1} - \mu = \rho(X_t - \mu) + \varepsilon_t$$

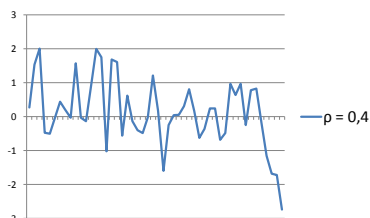
- Ofta men inte alltid fimpas man lägesparametern genom att istället titta på $Y_t = X_t - \mu$

$$Y_{t+1} = \rho Y_t + \varepsilon_t$$

Stokastiska processer 2



Mjukare, små förändringarna, slumpfelet mindre



"Taggigare", hoppar mer, slumpfelet slår igenom mer

Enkla index 1

- Välj en bastidpunkt ($t = 0$)
- Jämför samtliga observationer mot denna tidpunkt

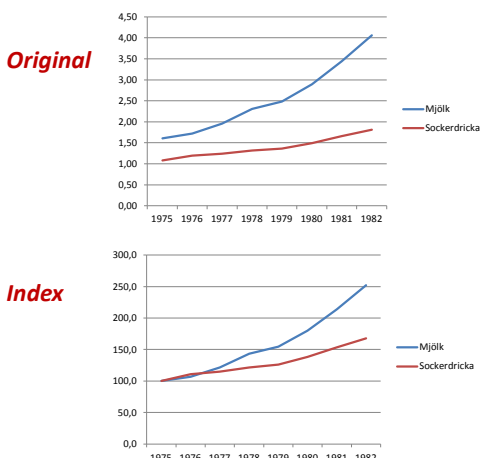
$$Index_t = 100 \cdot \frac{X_t}{X_0}$$

- Visar endast förändring, inte nivåer
- Endast jämföra mot basåret

År	X_t	$Index_t$
1975	1,61	100,0
1976	1,72	106,7
1977	1,96	121,6
1978	2,30	143,3
1979	2,48	154,5
1980	2,89	179,9
1981	3,44	214,2
1982	4,06	252,2

Enkla index 2

- I grafisk form:



Sammansatta index 1

- När man vill mäta t.ex. den allmänna prisnivån och dess förändringar (*inflation*).
- Ska endast mäta prisnivåns förändringar och inte påverkas av förändringar i konsumentens levnadsnivå.
- *Ett prisindex ska mäta prisnivåns förändringar vid oförändrad levnadsnivå.*

Sammansatta index 2

- Ex. priser på några varor:

	1981	1982
Kostym	845	932
Skjorta	102	117
Herrsockor	17	19

- Förändring?
 - Summera priserna och jämför summorna?
 - Jämför varje vara för sig och ta aritmetiskt medelvärde av förändringarna?

Sammansatta index 3

- Väg varje vara mot kvantiteterna, dvs. antalet sålda av var och en

$$Index_t = 100 \cdot \frac{\sum_{Varor} P_t Q}{\sum_{Varor} P_0 Q}$$

- Vilka kvantiteter ska man välja?
- År = 0 eller år t ?

Sammanfattning index 4

- **Laspeyres** index:

- Kvantiteter vid basåret = 0

$$Index_t^L = 100 \cdot \frac{\sum_{Varor} P_t Q_0}{\sum_{Varor} P_0 Q_0}$$

- **Paasche** index:

- Kvantiteter innevarande år t

$$Index_t^P = 100 \cdot \frac{\sum_{Varor} P_t Q_t}{\sum_{Varor} P_0 Q_t}$$

Sammanfattning index 5

- **Edgeworth** index:

- Snittet av kvantiteter vid basåret = 0 och t

$$Index_t^E = 100 \cdot \frac{\sum_{Varor} P_t \bar{Q}_{0,t}}{\sum_{Varor} P_0 \bar{Q}_{0,t}}$$

- **Fishers** idealindex:

- Geometriskt medelvärde av Laspeyres och Paasches

$$Index_t = \sqrt{Index_t^L \cdot Index_t^P}$$