Statistik 1 för biologer, logopeder och psykologer

Föreläsningar, del 1

Innehåll

Innehåll

Vad är statistik?

Statistik är inte läran om att föra statistik.

Statistiska undersökningar – målsättningar.

- Strävar i allmänhet efter att
  - beskriva
  - förklara
  - göra prognoser för
  - kontrollera

  olika fenomen.
- En förutsättning är att man ska kunna samla in information om fenomenet numerisk form.
- Man hoppas kunna skilja åt de *regelbundna* och de *slumpmässiga* karaktärsdragen i fenomenet.

Statistiska undersökningar – delmoment.

I en statistisk undersökning ingår i allmänhet följande tre moment:

- att insamla
- att sammanställa
- att dra slutsatser

av ett datamaterial.

Innehåll

Variabler

- En variabel är en storhet som varierar från individ till individ.
- Exempel på variabler är
  - längd
  - antal barn
  - utbildning
- En variabel kan anta olika värden.
- Ett datamaterial består av flera observationer på en eller flera variabler.

Kvantitativa och kvalitativa variabler

Kvantitativa variabler antar numeriska värden

- t.ex. ålder och vikt.

Kvalitativa variabler antar inte numeriska värden

- t.ex. utbildning och kön.

Kontinuerliga och diskreta variabler

Kontinuerliga variabler kan anta alla tänkbara värden i ett visst intervall

- t.ex. längd och vikt.

Diskreta kan endast anta vissa distinkta värden

- t.ex. antal barn och kön.

Skaltyper

Mätningar av variabler kan göras på olika skalnivåer:

- nominalskala
- ordinalskala
- intervallskala
- kvotskala

Skaltypen påverkar sättet att framställa och analysera datamaterialet.

Nominalskala

- Talar om för oss *vilken* klass en observation tillhör.
- Klasserna kan inte rangordnas sinsemellan.
- Exempel på observationer mätta på nominalskala:
  - kön
  - blodgrupp
  - hemstad.

Ordinalskala

- Talar om för oss vilken klass en observation tillhör samt om observationen har *mer* av en egenskap än en annan observation.
- Klasserna *kan* rangordnas sinsemellan
- Exempel på observationer mätta på ordinalskala:
  - militärgrad
  - klädstorlek: S, M, L, XL
  - vitsord i studentexamen.

Intervallskala

- Talar om för oss *hur mycket* en observation skiljer sig från en annan observation.
- Observationerna har numeriska värden men saknar en absolut nollpunkt.
- Exempel på observationer mätta på intervallskala:
  - temperatur mätt i Celsius eller Farenheit
  - vattenstånd i cm över en viss referenspunkt
  - datum.

Kvotskala

- Talar om för oss *hur många gånger* en observation har mer av en egenskap än en annan observation har.
- Numeriska värden som det är möjligt att bilda kvoter av.
- Exempel på observationer mätta på kvotskala:
  - temperatur mätt i Kelvin (där en absolut nollpunkt är definierad)
  - längd, bredd, vikt
  - ålder.

Fördelning av ett datamaterial

- För vidare analyser är det viktigt att bilda sig en uppfattning om det insamlade datamaterialet.
- Kan vara svårt att greppa utan någon form av sammanställning eller sammanfattning.
- Vi är ofta intresserade av hur observationerna är fördelade.
- Olika sätt att beskriva en fördelning:
  - tabulering
  - grafisk beskrivning
  - användning an central- och spridningsmått.

Frekvenstabeller

- Fördelningen av ett diskret datamaterial kan presenteras i form av en frekvenstabell
- Anger i tabellform antalet (=frekvensen) observationer tillhörande respektive klasser.
- Ofta anges också de relativa frekvenserna.

Frekvenstabeller

Exempel på en frekvenstabell.

åsikt	frekvens	rel. frekv.
A	12	$12/73 \approx 0.16$
B	29	$29/73 \approx 0.40$
C	14	$14/73 \approx 0.19$
D	7	$7/73 \approx 0.10$
E	11	$11/73 \approx 0.15$
sammanlagt	73	

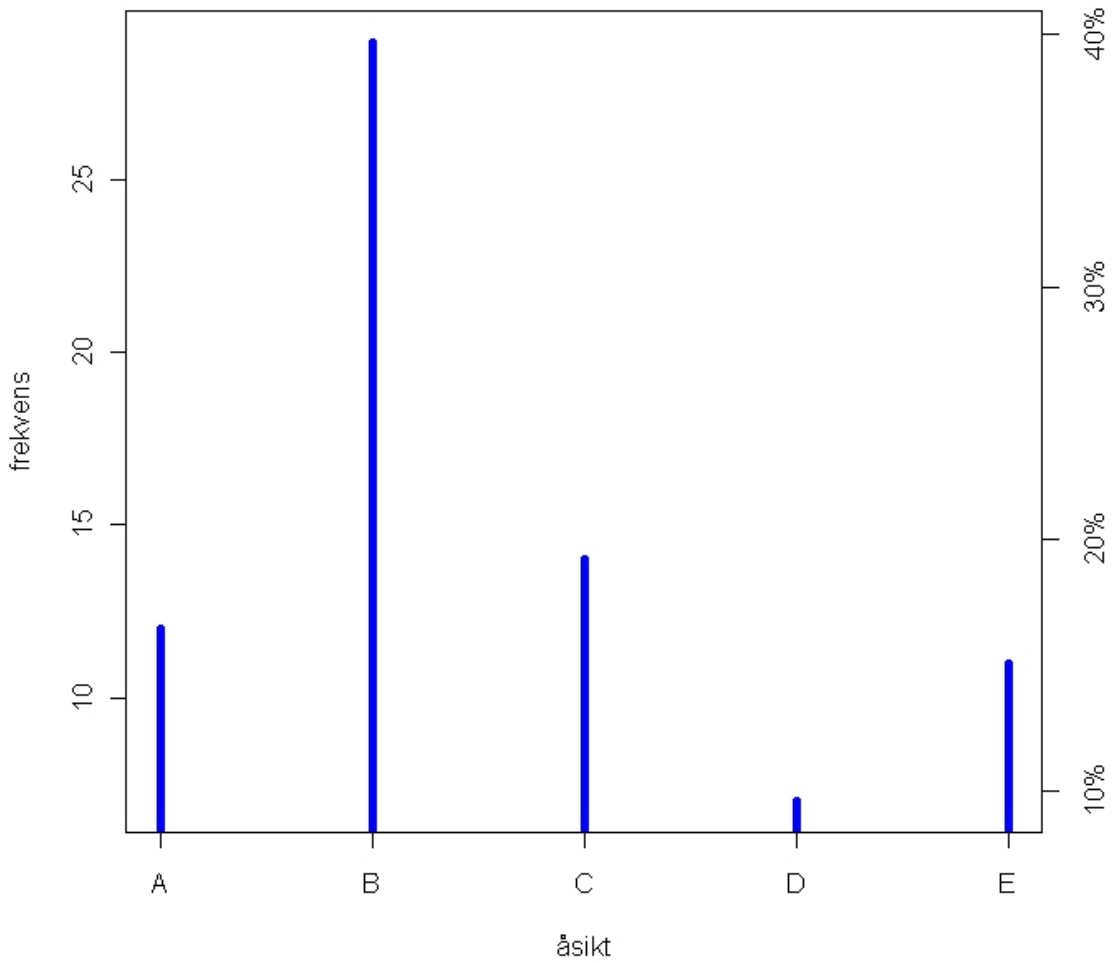
- De relativa frekvenserna kan också anges som procentandelar.

Stolpdiagram

- Diskreta fördelningar kan också presenteras grafiskt i form av stolpdiagram.
- Innehåller samma information som en frekvenstabell.
- En variant av stapeldiagram.

Stolpdiagram

Exempel på ett solpdiagram.



Korstabeller

Fördelningen av *bivariata* observarioner (två variabler observerade av samma individ) kan presenteras i form av en korstabell.

		INTELLIGENS		
		låg	medel	hög
ANPASSNINGS-FÖRMÅGA	låg	26	43	20
	medel	54	96	45
	hög	22	45	24

Klassindelning

- För kontinuerliga variabler får vi sällan två observationer med exakt samma värde.
- Det blir därmed meningslöst att i tabellform räkna upp frekvenserna för alla observerade värden.
- För att kunna konstruera en frekvenstabell blir det nödvändigt med klassindelning av datamaterialet.

Klassindelning

Frekvenstabell på klassindelat datamaterial över längder.

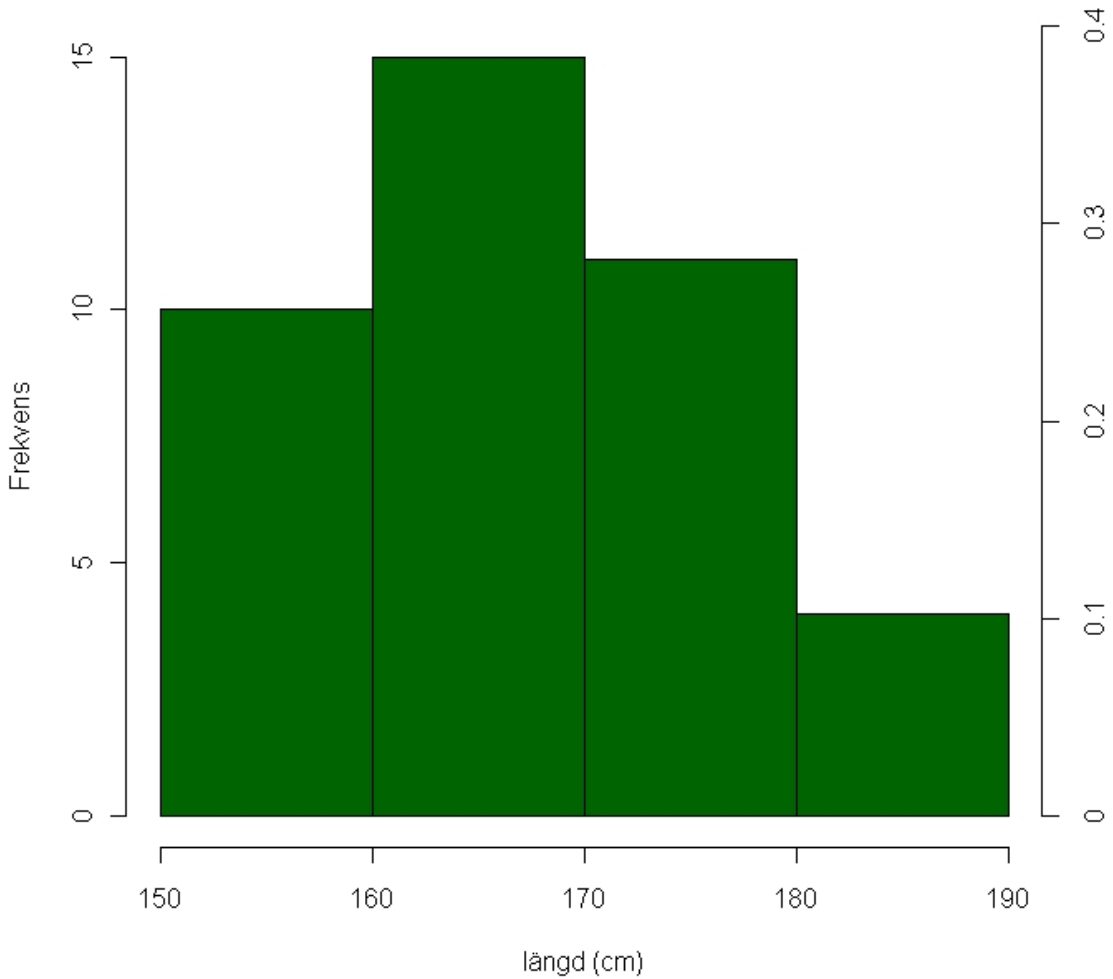
längd (cm)	frekvens	rel. frekv.
150–159	10	$10/39 \approx 0.26$
160–169	15	$15/39 \approx 0.38$
170–179	11	$11/39 \approx 0.28$
180–189	4	$4/39 \approx 0.10$
sammanlagt	39	

Histogram

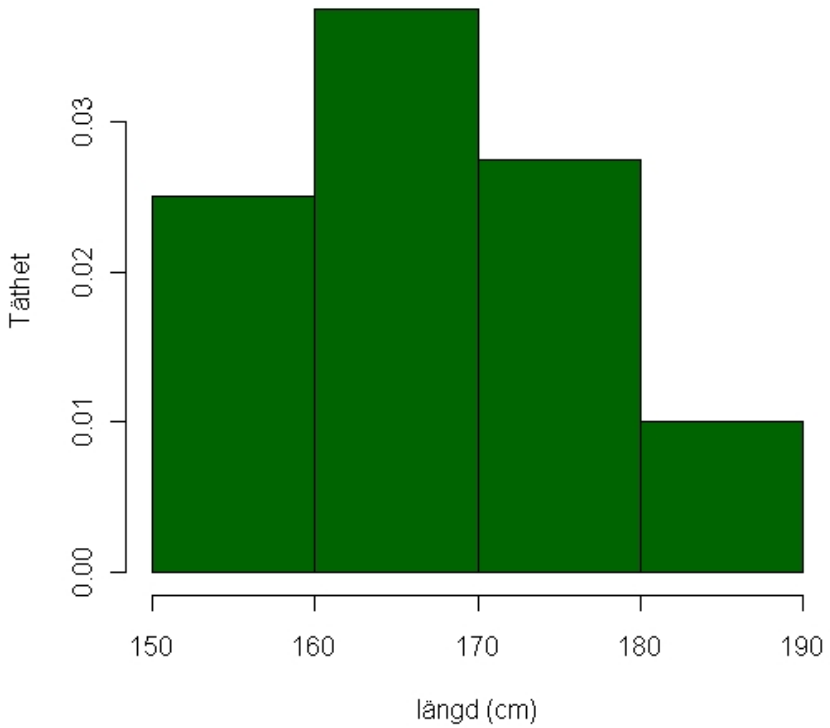
- Ett histogram är ett slags stapeldiagram där staplarna är fast i varandra.
- Om klasserna är lika breda motsvarar staplarnas höjd klassfrekvenserna.
- Om klasserna *inte* är lika breda måste frekvenserna korrigeras i förhållande till klassbredden.

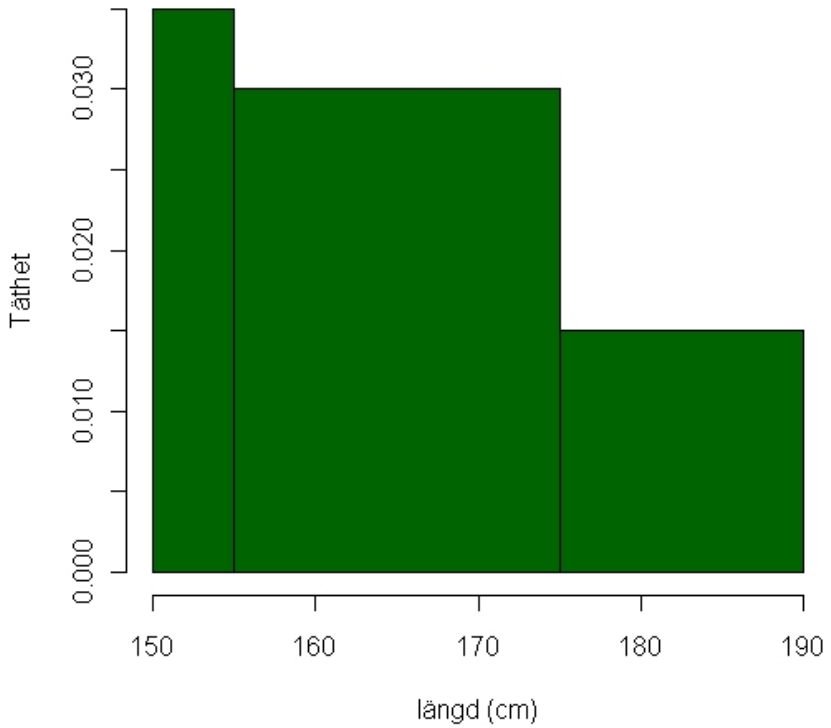
Histogram

Histogram på datamaterialet över längder.



Histogram över samma datamaterial med jämn och ojämn klassindelning.





Korrigering av klassfrekvens

klass	klassbredd	frekv.	korrigerad frekv.
12.5–13.4	1	2	3
13.5–14.4	1	7	7
14.5–15.4	2	12	$12/2 = 6$
16.5–20.4	4	6	$6/4 = 1.5$
sammanlagt		28	

Kumulativ frekvens

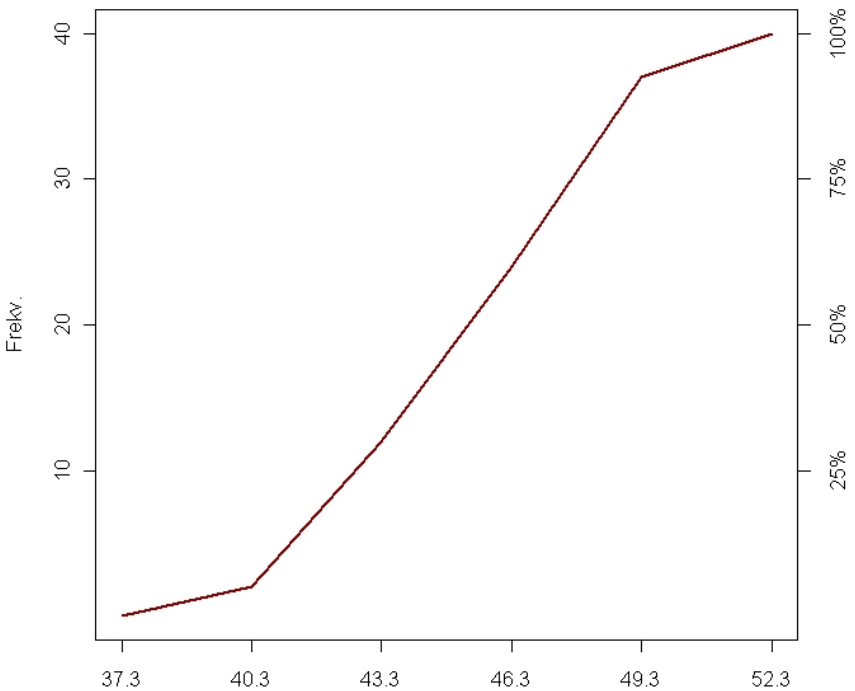
- Ofta vill inte endast veta klassfrekvenserna utan även hur många observationer som är mindre än ett visst värde.
- Vi talar då om den kumulativa frekvensen
- Den kumulativa frekvensen av den första klassen är samma som klassfrekvensen.

Kumulativ frekvens

klass	frekv.	kumulativ frekv.	relativ kumulativ frekv.
37.3–40.2	2	2	$2/40 \cdot 100\% = 5.0\%$
40.3–43.2	10	$2 + 10 = 12$	$12/40 \cdot 100\% = 30.0\%$
43.3–46.2	12	$12 + 12 = 24$	$24/40 \cdot 100\% = 60.0\%$
46.3–49.2	13	$24 + 13 = 37$	$37/40 \cdot 100\% = 92.5\%$
49.3–52.2	3	$37 + 3 = 40$	100%

Summapolygon

- Summaplygonet är en grafisk beskrivning av den kumulativa frekvensen.
- En annan benämning på summapolygonet är empirisk fördelningsfunktion.



Andra typer av diagram

- En förteckning över vanliga diagramtyper:
  `http://sv.wikipedia.org/wiki/Diagram`
- Se även den engelskspråkiga sidan för lite utförligare förklaringar av de vanligaste typerna:
  `http://en.wikipedia.org/wiki/Chart`

Innehåll

Typvärde

- Typvärdet är den klass / det värde som har den högsta frekvesnsen i en frekvenstabell.
- I ett stapeldiagram/histogram är den högsta stapeln vid typvärdet.
- Det kan finnas flera typvärden i ett datamaterial.
- Kan bestämmas för observationer mätta på alla skalnivåer.

Medelvärde

- Det aritmetiska melevärdet definieras som

$$\overline{x} = \frac{\text{summan av observationerna}}{\text{antalet observationer}} = \frac{\sum_{i=1}^{n} x_i}{n} .$$

- Medelvärdet kan endast räknas för intervall- och kvotskalevariabler.

Medelvärde

Exempel.

Medelvärdet av talen 3, 6 , 8, 4, 7, 1 räknas som

$$\frac{3+6+8+4+7+1}{5} = 4.83 \ .$$

Viktat medelvärde

- För att räkna medelvärdet på ett datamaterial på basen av en frekvenstabell använder vi oss av ett viktat medelvärdet.
- Det viktade medelvärdet räknas genom att vikta (=multiplicera) variabelns varje värde med antalet gånger det har dykt upp i datamaterialet (=frekvensen).
- För ett kontinuerligt klassindelat datamaterial kan klassmedelpunkterna användas som värde för variabeln.

Viktat medelvärde

Exempel.

Vi betraktar följande frekvenstabell:

antal bilar per hushåll	frekvens
0	5
1	3
2	1
3	1
sammanlagt	10

Det viktade medlvärdet av datamaterialet är

$$\frac{5 \cdot 0 + 3 \cdot 1 + 1 \cdot 2 + 1 \cdot 3}{10} = 0.8 \, .$$

Viktat medelvärde med relativa frekvenser.

Har vi tillgång till de relativa frekvenserna får vi medeltalet direkt som en viktad *summa* av de observerade värdena.

Exempel.

antal bilar per hushåll	frekvens	rel. frekv.
0	5	5/10=0.5
1	3	3/10=0.3
2	1	1/10=0.1
3	1	1/10=0.1
sammanlagt	10	

Det viktade medlvärdet räknas nu som

$$0.5 \cdot 0 + 0.3 \cdot 1 + 0.2 \cdot 1 + 0.2 \cdot 1 = 0.8 \ .$$

Jfr uträkningen på föregående sida!

Median

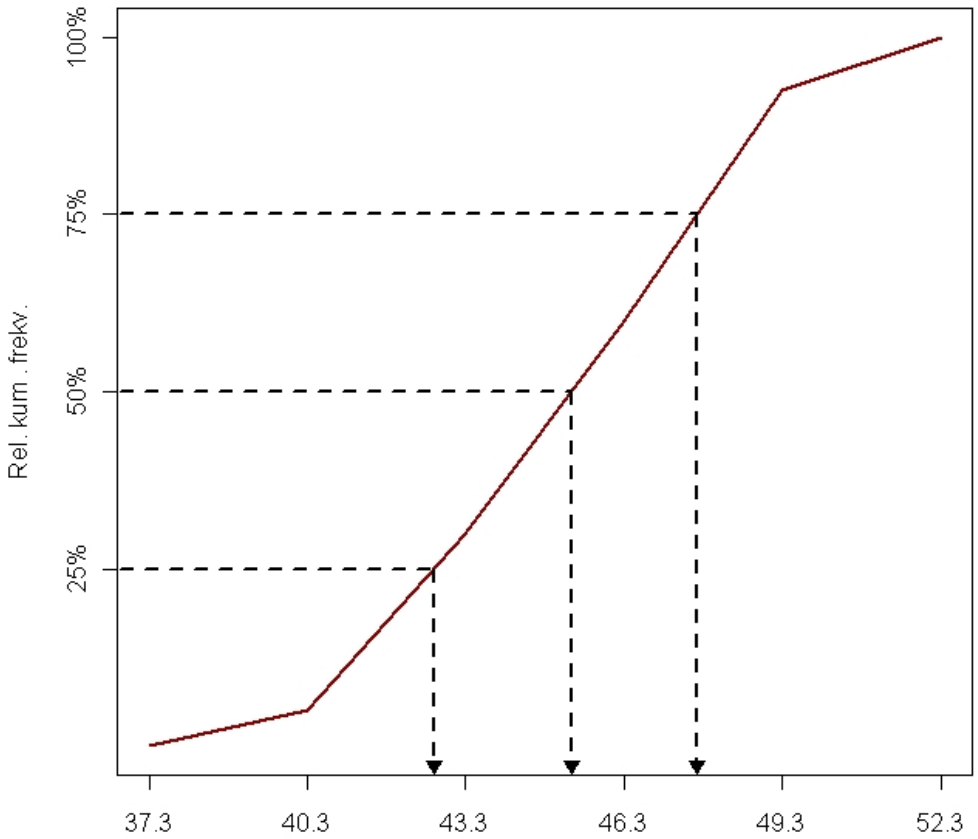
- Medianen är det mittersta värdet i ett datamaterial som är ordnat från den minsta observationen till den största.
- Medianen lämpar sig för variabler av alla andra skaltyper förutom nominalskalan.
- Om datamaterialet innehåller ett jämnt antal observationer är medianen någondera av de två mittersta observationerna eller, om möjligt, medeltalet av dem.

Median, kvantiler och fraktiler

- För ett kontinuerligt datamaterial är medianen det värde där den relativa kumulativa frekvensen är 0.5 (dvs. 50%).
- På samma sätt kan även den undre kvartilen (rel. kum. frekv. 0.25) och övre kvartilen (rel. kum. frekv. 0.75) bestämmas.
- Mer allmänt kan fraktiler för vilken relativ kumulativ frekvens som helst betämmas.
- Om de relativa kumulativa frekvenserna är angivna i procent kallas fraktilerna ofta *percentiler*.

Grafisk bestämning av median och kvartiler

Medianen, kvantiler och fraktiler kan även bestämmas grafiskt från en empirisk fördelningsfunktion.



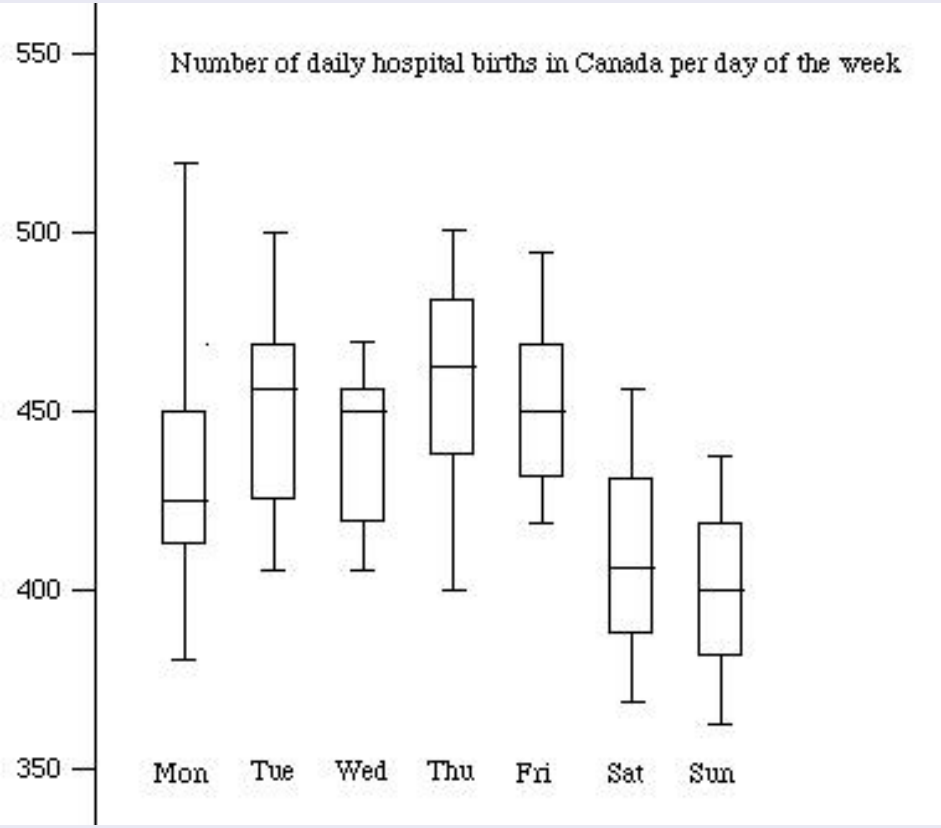
Boxplot

En boxplot (även kallad lådogram) sammanfattar grafist en fördelning i följande 5 punkter (räknat uppifrån ner):

- högsta värdet
- övre kvartilen
- medianen
- undre kvartilen
- mista värdet.

Boxplotten lämpar sig speciellt bra för jämförelser av samma variabel över olika grupper eller experiment.

Exempel.



Variationsvidd

- Variationsvidden anger avståndet mellan det mista och det största värdet i ett datamaterial.
- Variationsvidden definieras som

$$\text{variationsvidd} = \text{största värdet} - \text{minsta värdet}\ .$$

- Variationsvidden kan endast bestämmas för intervall- och kvotskalevariabler.

Kvartilavstånd

- Kvartilavståndet anger avståndet mellan den undre och den övre kvartilen.
- Kvartilavståndet definieras som

$$\text{kvartilavstånd} = \text{övre kvartilen} - \text{undre kvartilen}\ .$$

- Kvartilavståndet kan endast bestämmas för intervall- och kvotskalevariabler.

Standardavvikelse och varians.

- Standardavvikelsen anger hur mycket observationerna i ett kvantitativt datamaterial i *genomsnitt* avviker från medelvärdet.
- Standardavvikelsen beräknas enligt

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

där x_i är den i:te observationen, $\overline{x}$ är medelvärdet på alla observationer och n är antalet observationer.

- Lämnar vi bort kvadratroten i uttrycket ovan får vi variansen s^2.
- Standardavvikelsen anges i samma enhet som observationerna (t.ex. *cm*) medan variansen är en dimensionslös storhet.

Standardavvikelse och varians.

Exempel.

Vi räknar variansen och standardavvikelsen av talen

$$1, 3, 5, 8, 9, 10 \ .$$

Vi börjar med att räkna avvikelsen mellan varje observation och deras medelvärde 6. Då avvikelserna är

$$-5, -3, -1, 2, 3, 4$$

och det totala antalet observationer är 6, blir variansen

$$\frac{(-5)^2 + (-3)^2 + (-1)^2 + 2^2 + 3^2 + 4^2}{5} = 12.8$$

och standardavvikelsen $\sqrt{12.8} \approx 3.6$.

Variationskoefficient

- Variationskoefficienten är en normaliserad standardavvikelse som uttrycker hur många procent i genomsnitt observationerna avviker från medelvärdet.
- Variationskoefficienten definieras som

$$\text{variationskoefficient} = \frac{\text{standardavvikelse}}{\text{medelvärde}} \cdot 100\% = \frac{s}{\overline{x}} \cdot 100\% \, .$$

- Gör standardavvikelser på datamaterial där observationerna är mätta i olika enheter jämförbara.
- Används endast på icke-negativa data.

Summatecken

- Kort om användning av summatecknet $\sum$ som förekommer i en del av formlerna för central- och spridningsmått: `http://sv.wikipedia.org/wiki/Summatecken`

Fallgropar med vanliga central- och spridningsmått

http://web.abo.fi/fak/mnf/mate/jc/statistik1/DeskriptivtExempel.pdf