Statistisk översiktskurs - Föreläsning 2

Anders Fredriksson

Statistiska Institutionen
Stockholms Universitet

March 24, 2025

Föreläsning 2

- Variabeltyper och variabelegenskaper
- Deskriptiv statistik: Olika typer av tabeller, diagram och grafer
- Sammanfattande mått (lägesmått och spridningsmått)

Variabler

- En variabel är, som ordet antyder, något som kan variera
- En egenskap som kan variera mellan olika element i populationen
- Delas in i kategoriska variabler och numeriska variabler
- I ett välstrukturerat dataset representerar varje kolumn en variabel

Kategoriska variabler

Kategoriska variabler (alternativt: Kvalitativa, Icke-numeriska)

- Kan delas in i kategorier, grupper

Exempel

- kön
- civilstånd
- bilmärken
- studieprogram
- partitillhörighet

Numeriska variabler

Numeriska variabler (alternativt: Kvantitativa)

- Representerar mängder

Exempel

- antal bilar på en parkering
- mängd flingor i ett paket (i gram)
- en individs längd (i cm)
- andel som röstar på ett visst parti (i procent)

Numeriska variabler kan vidare delas in i

Diskreta numeriska variabler

- kan endast anta vissa värden (oftast 0, 1, 2, osv.)
- ex: antal paket flingor

Kontinuerliga numeriska variabler

- kan anta alla värden
- ex: mängd flingor i ett paket (i gram)

Exempel på variabler och *variabeltyp*

- **Färgpreferens**
  Kategorisk
- **Antal hemmavarande barn**
  Kvantitativ och diskret
- **Tid att springa 100 meter**
  Kvantitativ och kontinuerlig
- **Attityd till återinförande av fastighetsskatt**
  Kategorisk
- **Poäng på tenta**
  Kvantitativ och i praktiken diskret
- **Partisympati**
  Kategorisk
- **Ålder**
  Kvantitativ och om "antal fyllda år" diskret

Datanivåer / skalnivåer

Utöver indelningen i kategorisk/numerisk delar man också in variabler efter s.k. skalnivå (ungefärligen: informationsinnehåll)

- **Nominalskala** - klassificerar men går inte att rangordna
  ex: färg - gul/grön/blå

- **Ordinalskala** - går att rangordna
  ex: gillar du regeringen? - lite/mellan/mycket

- **Intervallskala** - samma avstånd på skalan (”skalsteg”)
  ex: temperatur i Celsius

  (skillnaden mellan 3 och 2 Celsius är lika med
  skillnaden mellan 15 och 14 Celsius)

- **Kvotskala** - absolut nollpunkt
  ex: temperatur i Kelvin, antal barn, avstånd till olika turistmål

Skalnivåer, tabell

	Rangordning	**Lika skalsteg**	**Absolut 0-punkt**
Nominal	Nej	Nej	Nej
Ordinal	Ja	Nej	Nej
Intervall	Ja	Ja	Nej
Kvot	Ja	Ja	Ja

Tre saker att hålla reda på

- En kategorisk variabel är antingen på nominal- eller ordinalskala

- Även om vi kodar en kategorisk variabel med siffror innebär det inte att vi får en numerisk variabel

  Ex: civilstånd enligt 1=ogift, 2=gift, 3=..., osv.
  Siffrorna i sig har ingen kvantitativ innebörd

  Ex: Gillar du regeringen? (1=lite, 2=mellan, 3=mycket)

  Skillnad mellan 2 & 1, och mellan 3 & 2, måste inte vara samma

- En numerisk variabel kan göras om till en kategorisk variabel

  Ex: dela in ålder i ålderskategorier, exv. 18-30, 31-40, osv.

  (vi får en kategorisk variabel på ordinalskala)

Deskription - Att sammanfatta, beskriva och redovisa data.

Vi börjar med: Kategoriska variabler - frekvenstabell

En kategorisk variabel

- Ex: Titanic, överlevande
- Ex: Titanic, klass

Status_överlevnad	Antal_individer	Andel_individer	Procentandel_ind
Överlevde	712	0.322	32.2
Överlevde inte	1496	0.678	67.8
Totalt	2208	1.000	100.0

Klass	Antal_individer	Andel_individer	Procentandel_ind
Första	324	0.147	14.7
Andra	285	0.129	12.9
Tredje	710	0.322	32.2
Besättning	889	0.403	40.3
Totalt	2208	1.000	100.0

Kategoriska variabler - korstabell

Två kategoriska variabler

- Ex: Titanic, överlevande och klass
- Begreppen **simultan fördelning** och **marginell fördelning** (återkommer på datorlaboration 2)

Klass	Överlevde (%)	Överlevde inte (%)	Totalt (%)
Första	9.1	5.6	14.7
Andra	5.4	7.5	12.9
Tredje	8.2	24.0	32.2
Besättning	9.6	30.7	40.3
Totalt	32.2	67.8	100.0

Kategoriska variabler - korstabell med betingad fördelning

- Två kategoriska variabler
- Ex: Titanic, överlevande per klass ("betingat på klass")
- Begreppet **betingad fördelning** (återkommer på datorlaboration 2)

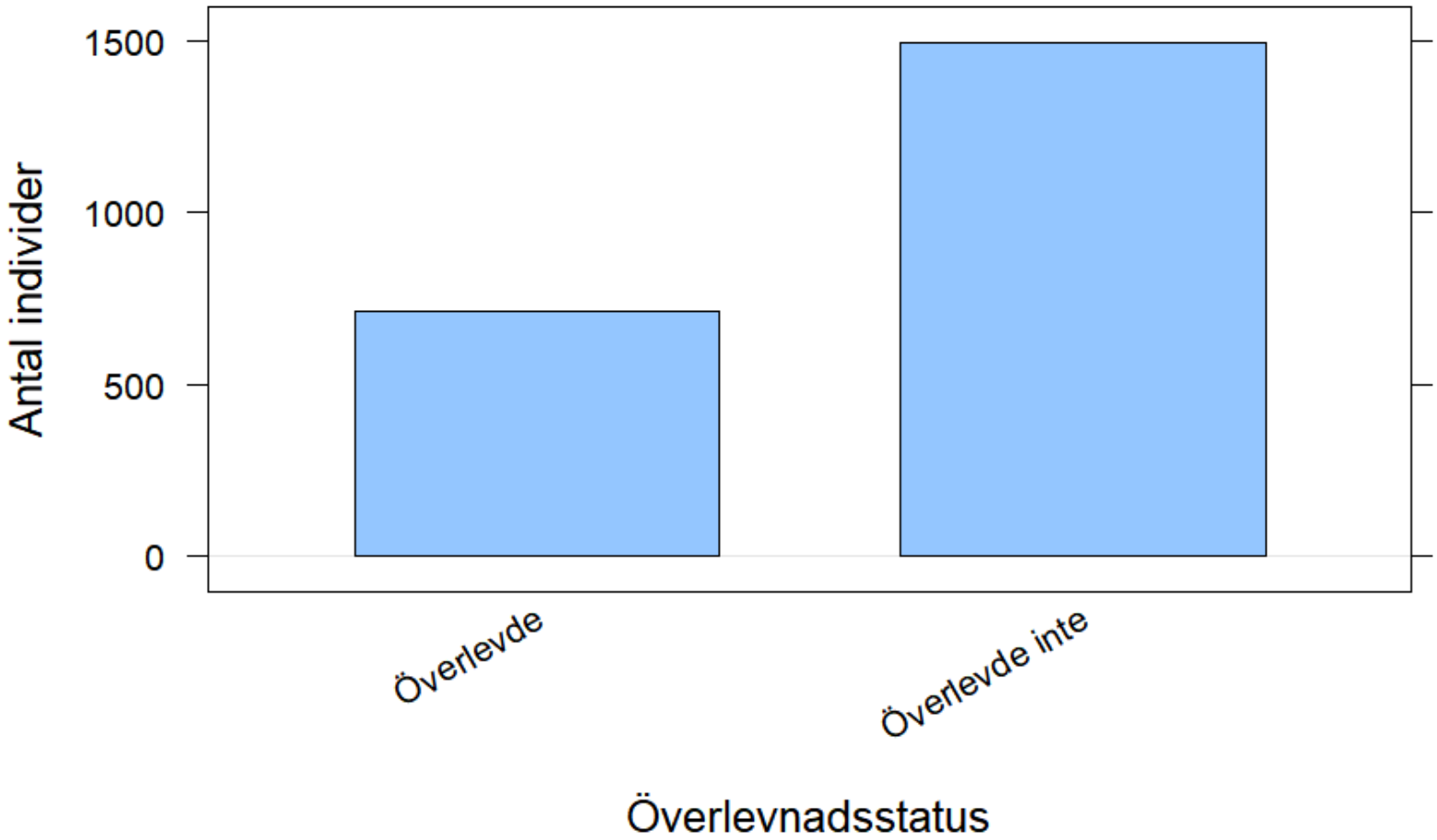
Status_överlevnad	Första kl.	Andra kl.	Tredje kl.	Besättning
Överlevde (%)	62	41.8	25.4	23.8
Överlevde inte (%)	38	58.2	74.6	76.2
Totalt (%)	100	100.0	100.0	100.0

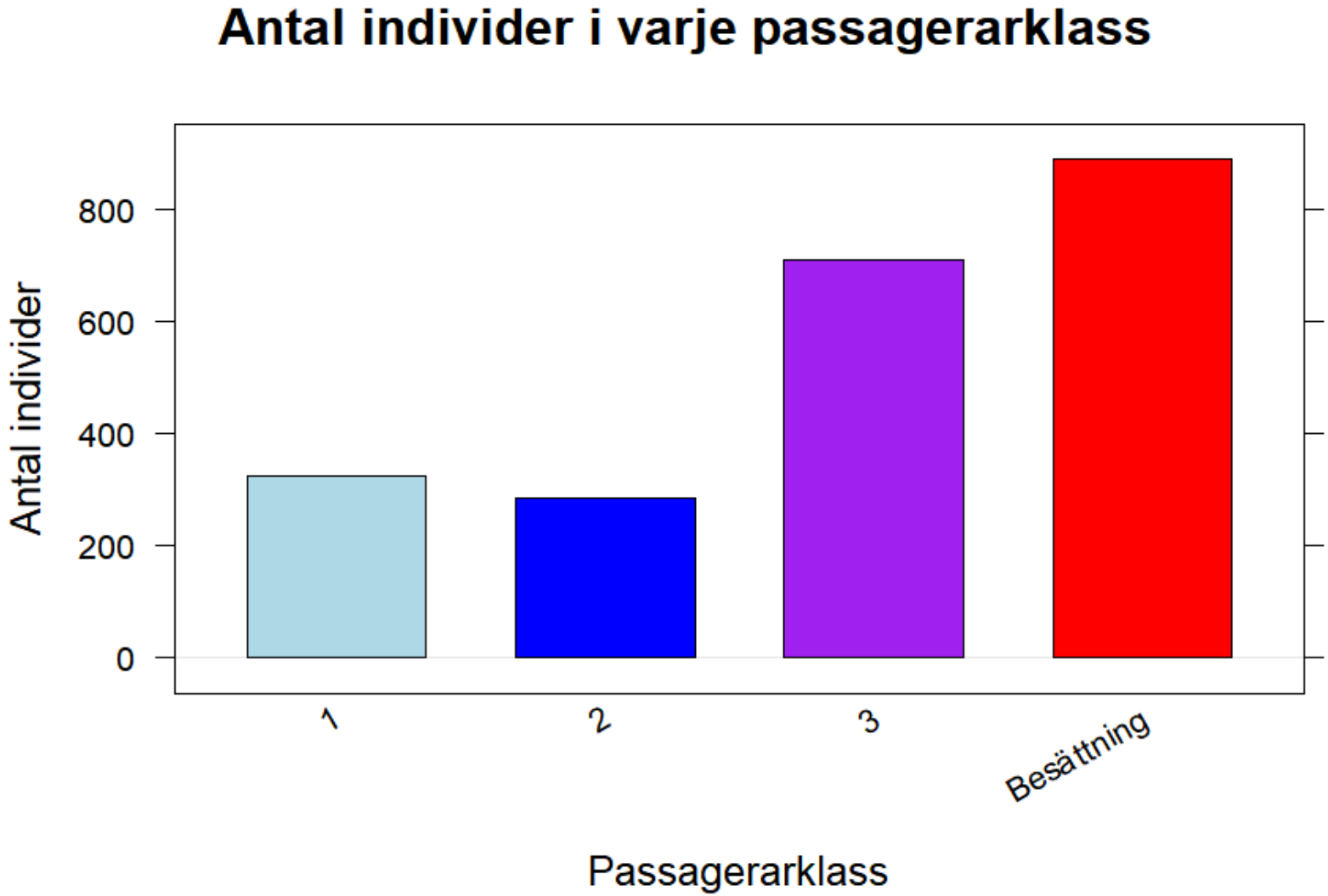
Kategoriska data - stapeldiagram (bar chart)

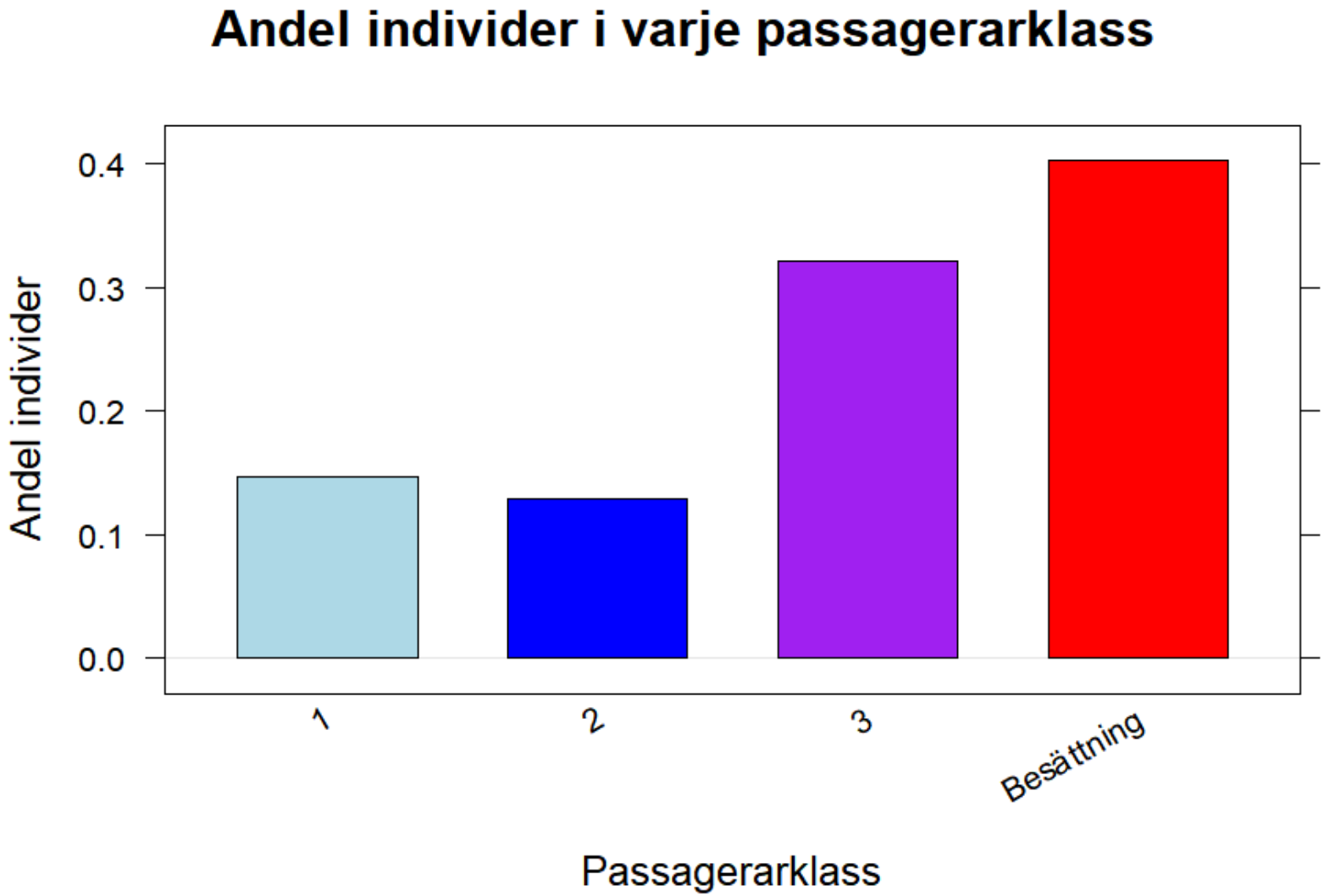
En kategorisk variabel

- Ex: Titanic, överlevande
- Ex: Titanic, klass

Antal individer som överlevde/inte överlevde Titanic




Antal individer i varje passagerarklass
Antal individer
0
200
400
600
800
1
2
3
Besättning
Passagerarklass

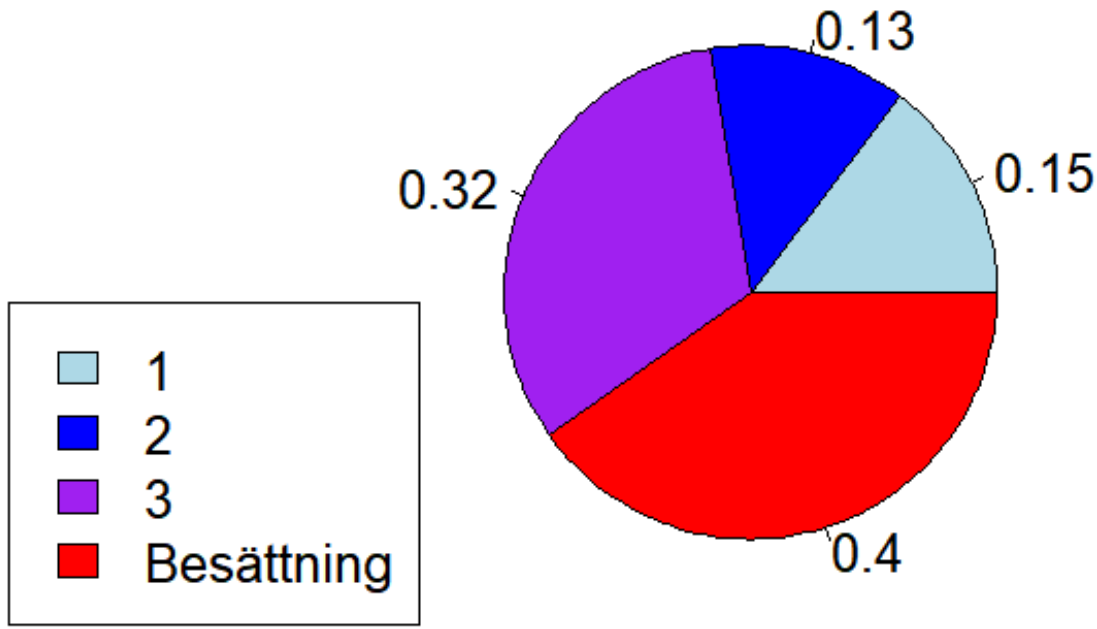

Andel individer i varje passagerarklass
Andel individer
0.0
0.1
0.2
0.3
0.4
1
2
3
Besättning
Passagerarklass

Kategoriska data - cirkeldiagram (pie chart)

En kategorisk variabel

- Ex: Titanic, klass

Andel individer i varje passagerarklass

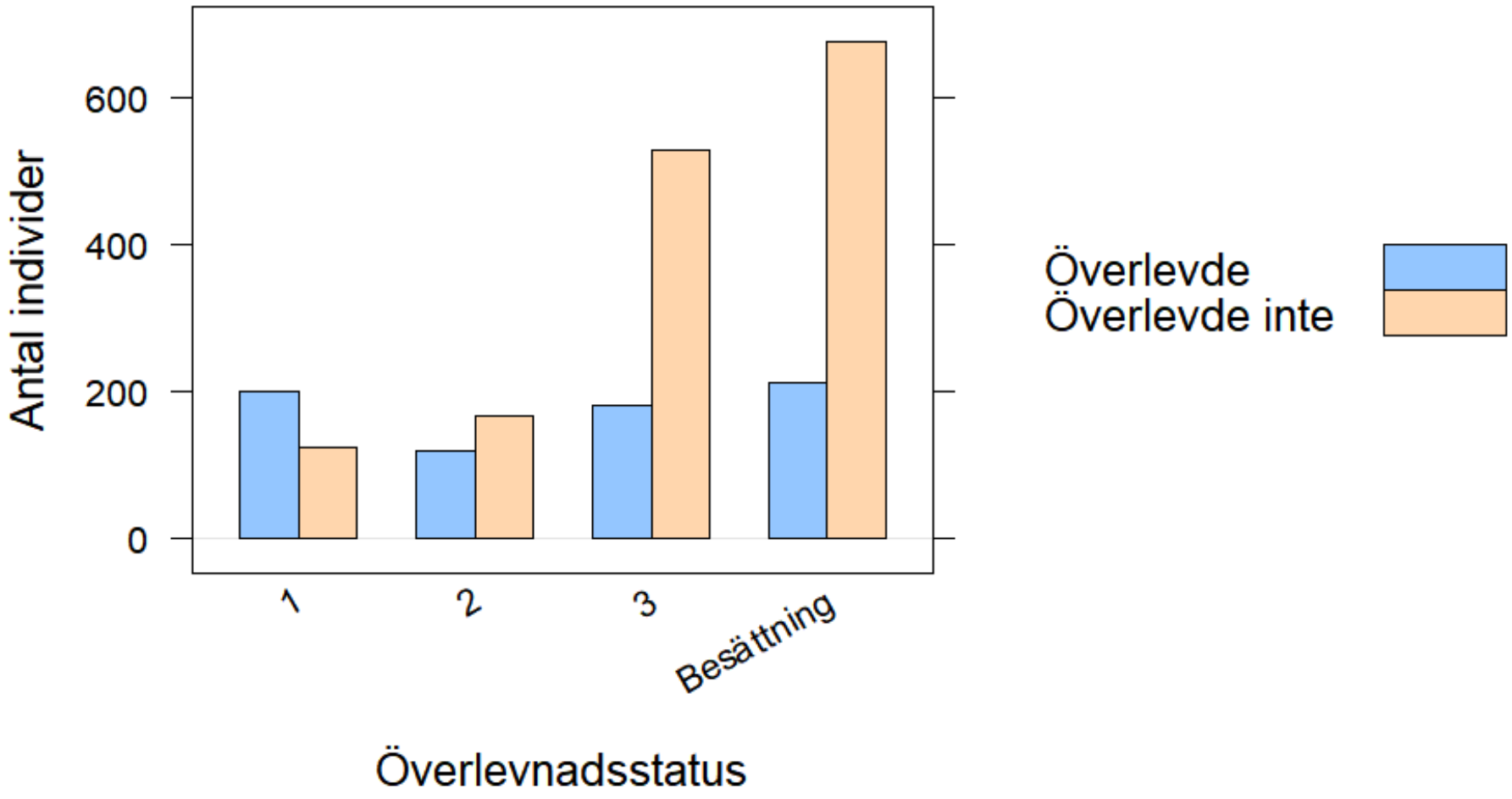


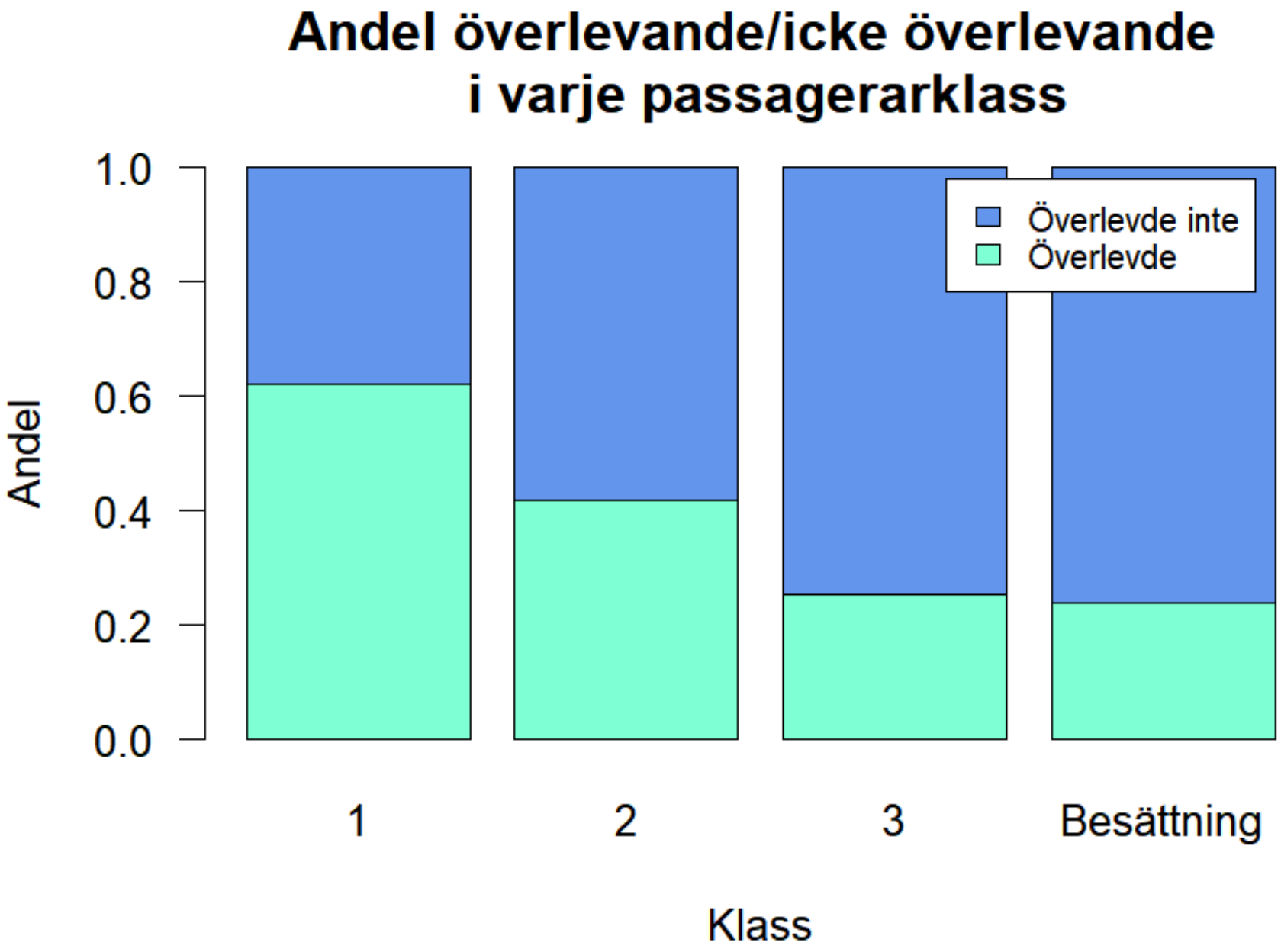
Kategoriska data - grupperade och staplade stapeldiagram (grouped / stacked bar charts)

Två kategoriska variabler

- Ex: Titanic, överlevande per klass

Antal överlevande/icke överlevande i varje passagerarklass



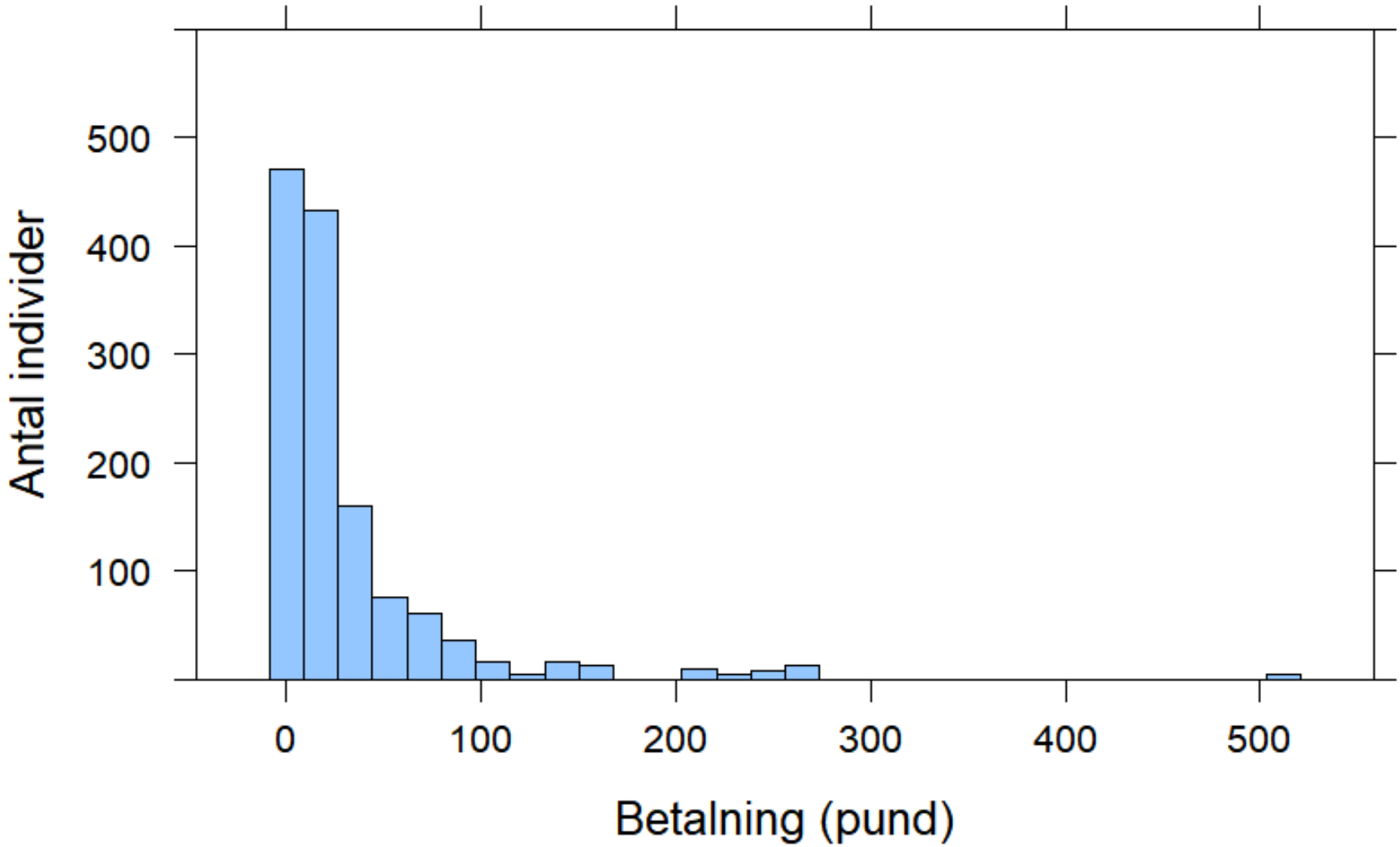

Andel överlevande/icke överlevande
i varje passagerarklass
1.0
0.8
0.6
0.4
0.2
0.0
Andel
Överlevde inte
Överlevde
1
2
3
Besättning
Klass

Numeriska data - histogram (histogram)

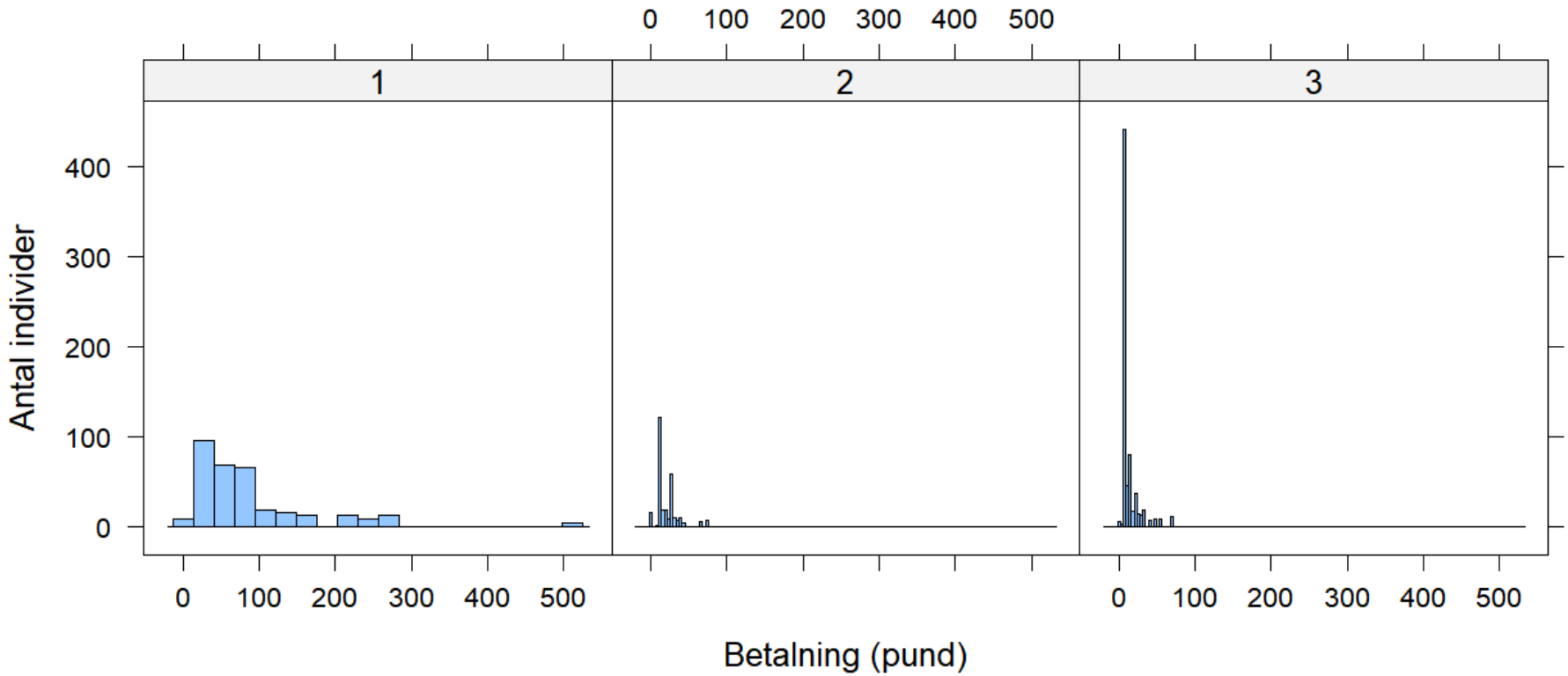
En numerisk variabel

- Ex: Titanic, biljettpris
- Skilj på histogram och stapeldiagram (som används för kategoriska variabler)
- I ett histogram kommer bredden på staplarna (hur många staplar vi har) spela roll för exakt vilken information vi får från figuren
- Mer om detta på datorövningarna

Fördelning av kostnaden för Titanicresan (för 1318 av 2208 passagerare)



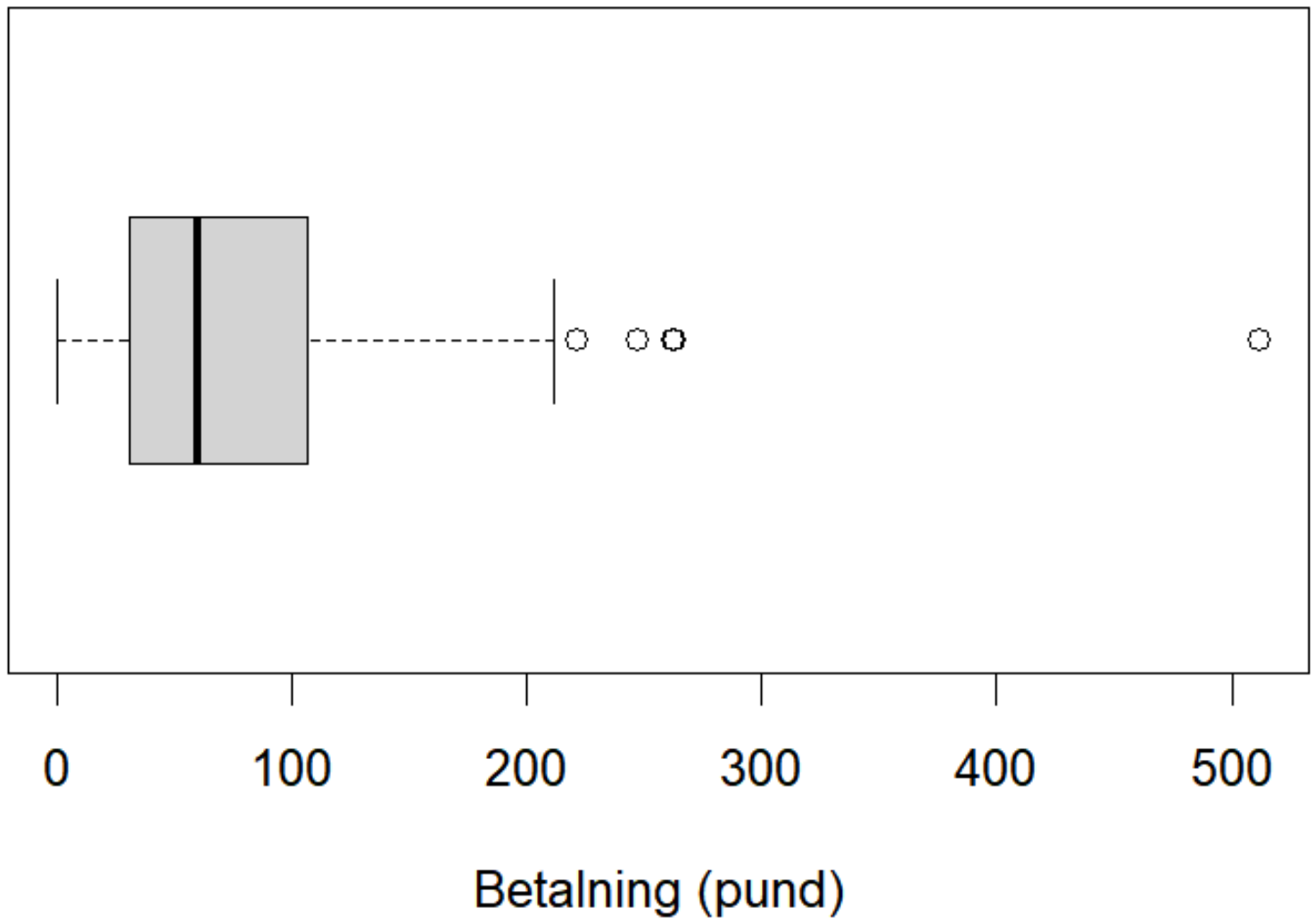
Fördelning av kostnaden för Titanicresan per klass (för 1318 av 2208 passagerare)



Numeriska data - lådagram/låddiagram (boxplot)

En numerisk variabel

- Ex: Titanic, biljettpris
- Mer detaljer kommer på laboration 2


Fördelning av kostnaden för Titanicresan, i klass 1
0
100
200
300
400
500
Betalning (pund)

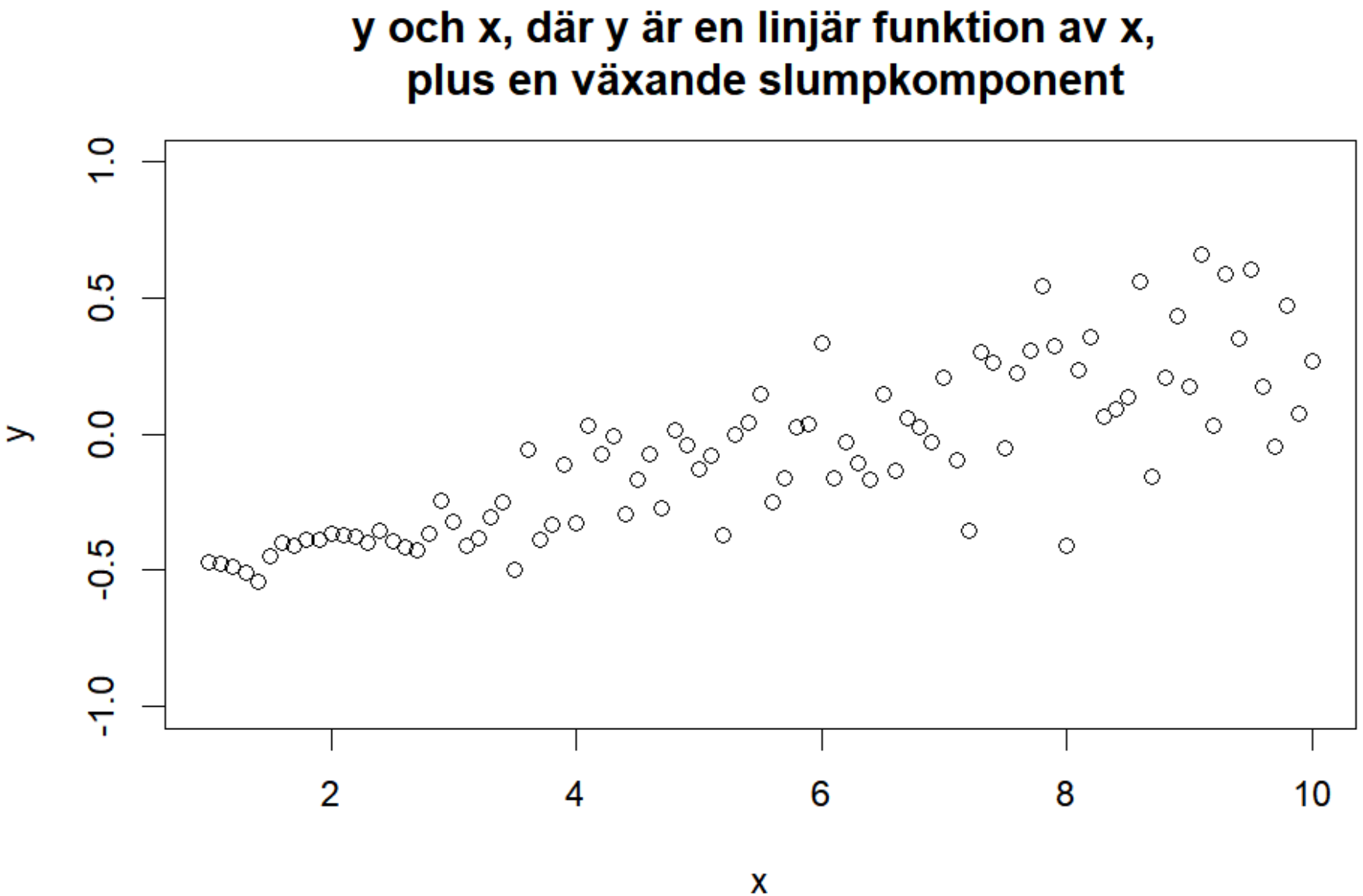
Numeriska data - spridningsdiagram (scatter plot)

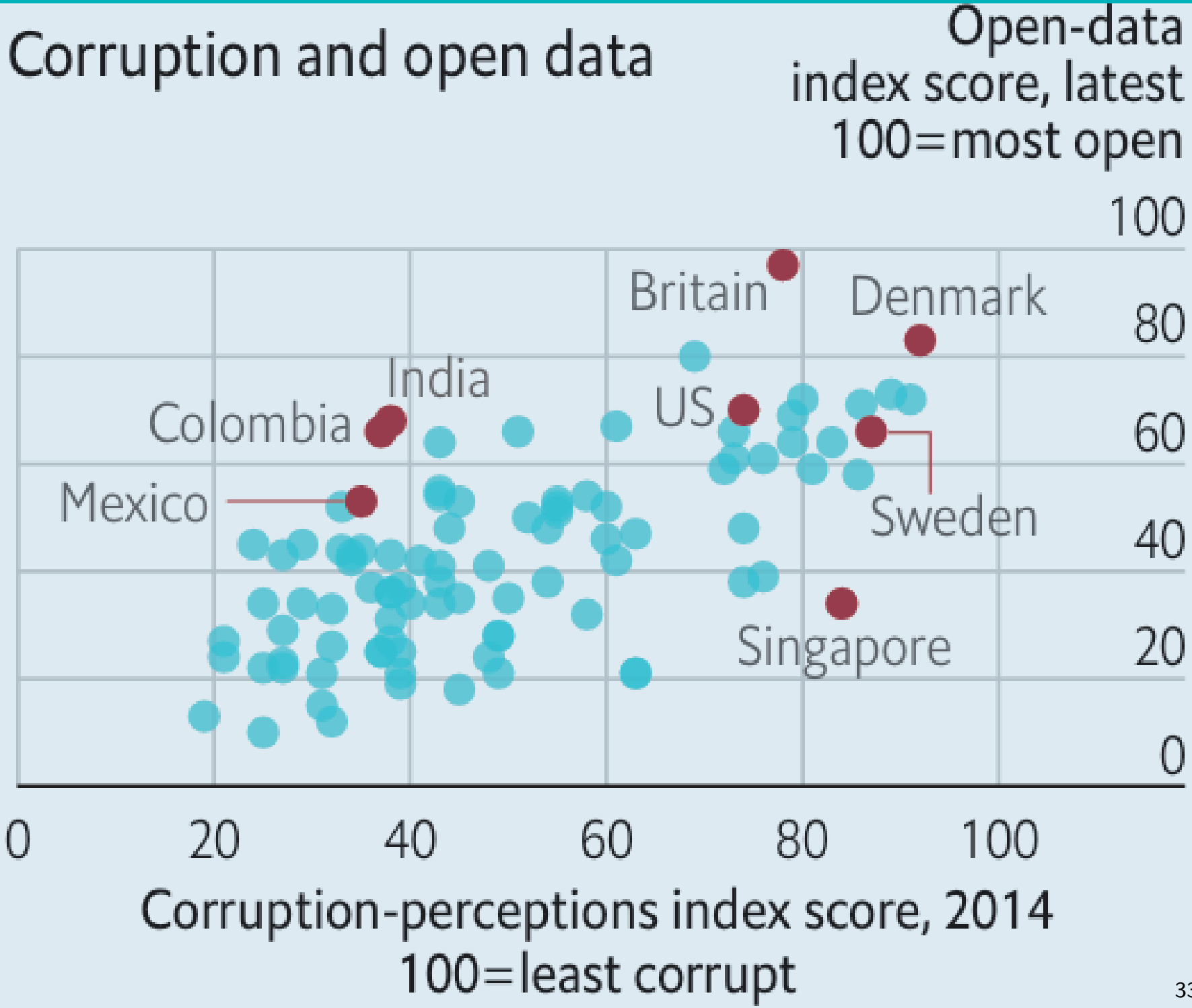
Två numeriska variabler

- Färg, storlek, etc., kan användas för att åskådliggöra ytterligare variabler (typiskt kategoriska)

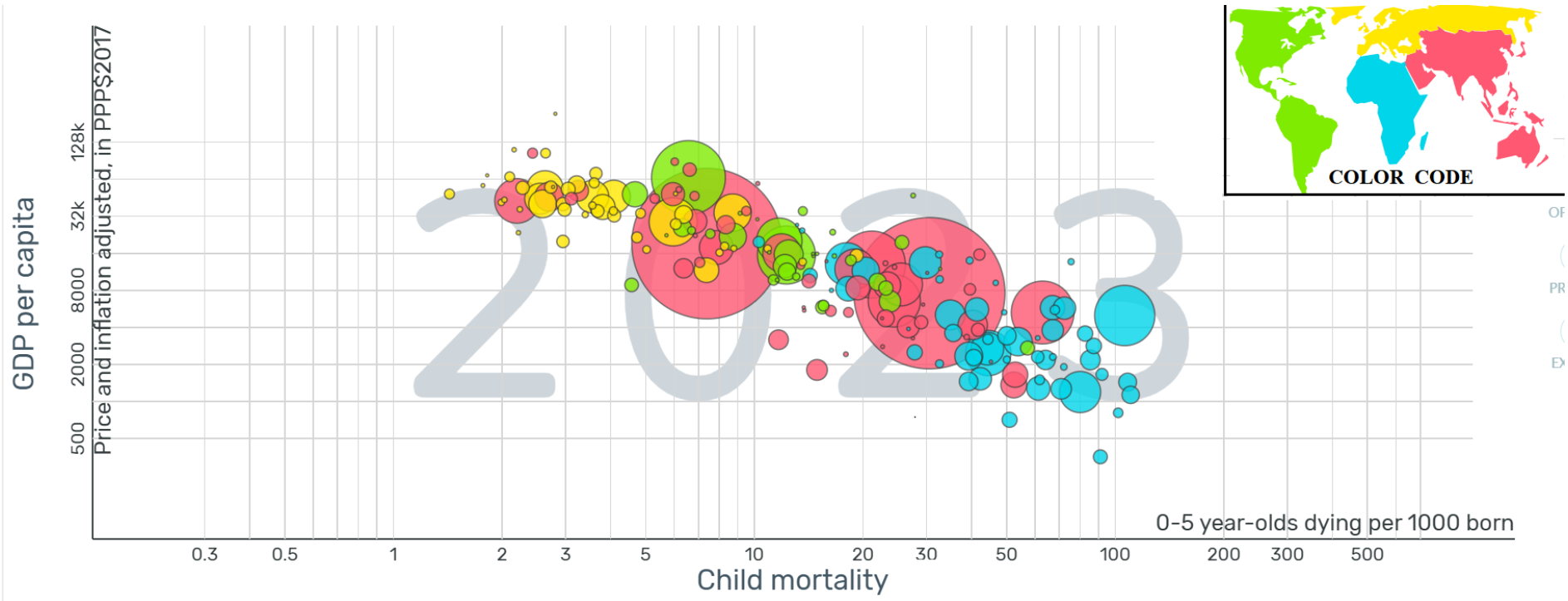
- Exempel med påhittade data
- Exempel från The Economist
- Exempel från Gapminder (Hans Roslings presentationer)

Exempel på spridningsdiagram - påhittade data




Corruption and open data
Open-data index score, latest 100=most open
100
80
60
40
20
0
Britain
Denmark
India
US
Colombia
Mexico
Sweden
Singapore
0
20
40
60
80
100
Corruption-perceptions index score, 2014
100=least corrupt

BNP per capita och barnadödlighet (storlek på cirkel - (relativ) folkmängd)



Källa: Gapminder

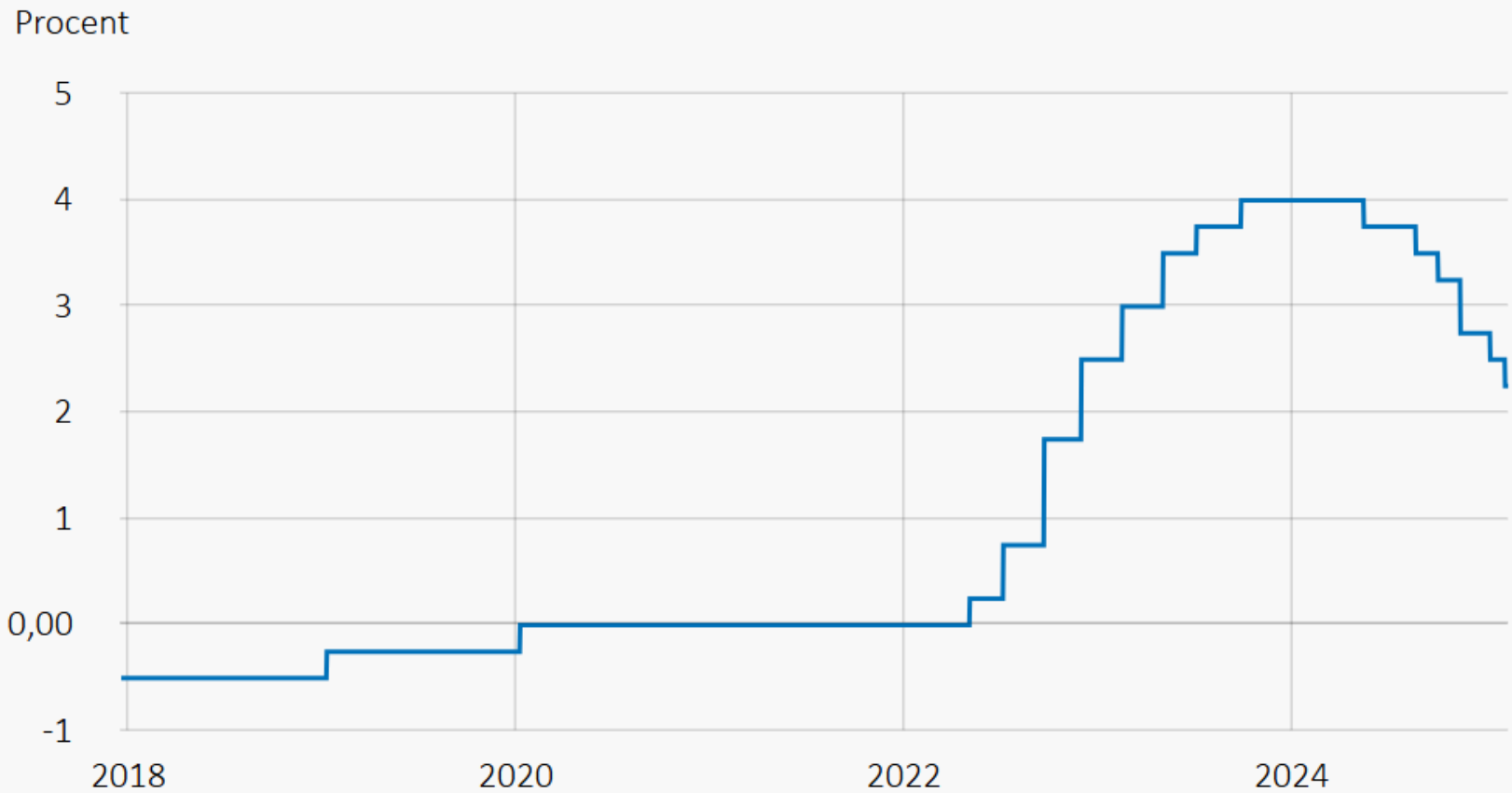
Numeriska data - tidsseriediagram (time series diagram)

En eller flera variabler som har en tidsdimension

- Exempel från Riksbanken

Styrräntan

Diagrammet visar styrräntan 2018–2024



Källa Riksbanken.

När ni gör figurer

- I alla grafer - ange vad som finns på axlarna, inklusive enhet
- Ha en informativ rubrik som kortfattat förklarar vad som visas
- Ha lämpliga intervall på axlarna, som inte vilseleder läsaren
- För akademiska artiklar - ha tabellrubrik ovanför tabellen och figurrubrik under figuren. I materialet här har vi alla rubriker ovanför.

Lägesmått: Typvärde, median och medelvärde

- Antag att vi har data på antalet barn i 9 familjer
- Kalla variabeln ”antal barn” för X (stora X)
- Kalla storleken på vårt dataset för n (9 i detta exempel)
- Indexera varje observation (varje familj) med bokstaven i
- $x_1, x_2, \ldots x_i \ldots, x_n$ (lilla x) är värden på olika observationer av X, i vårt dataset, för i mellan 1 och n

- Exempeldata på antal barn i 9 familjer
- $x_1, \ldots, x_9 = \{0, 3, 1, 8, 1, 2, 2, 0, 1\}$

Typvärde, median och medelvärde

Data igen: $x_1, \ldots, x_9 = \{0, 3, 1, 8, 1, 2, 2, 0, 1\}$

- **Typvärdet** ("mode") är det vanligaste värdet

  Typvärde $= 1$

- **Medianen** ("median") - rangordna värdena från minst till störst och ta fram det mittersta värdet

  $\{0, 0, 1, 1, \mathbf{1}, 2, 2, 3, 8\}$

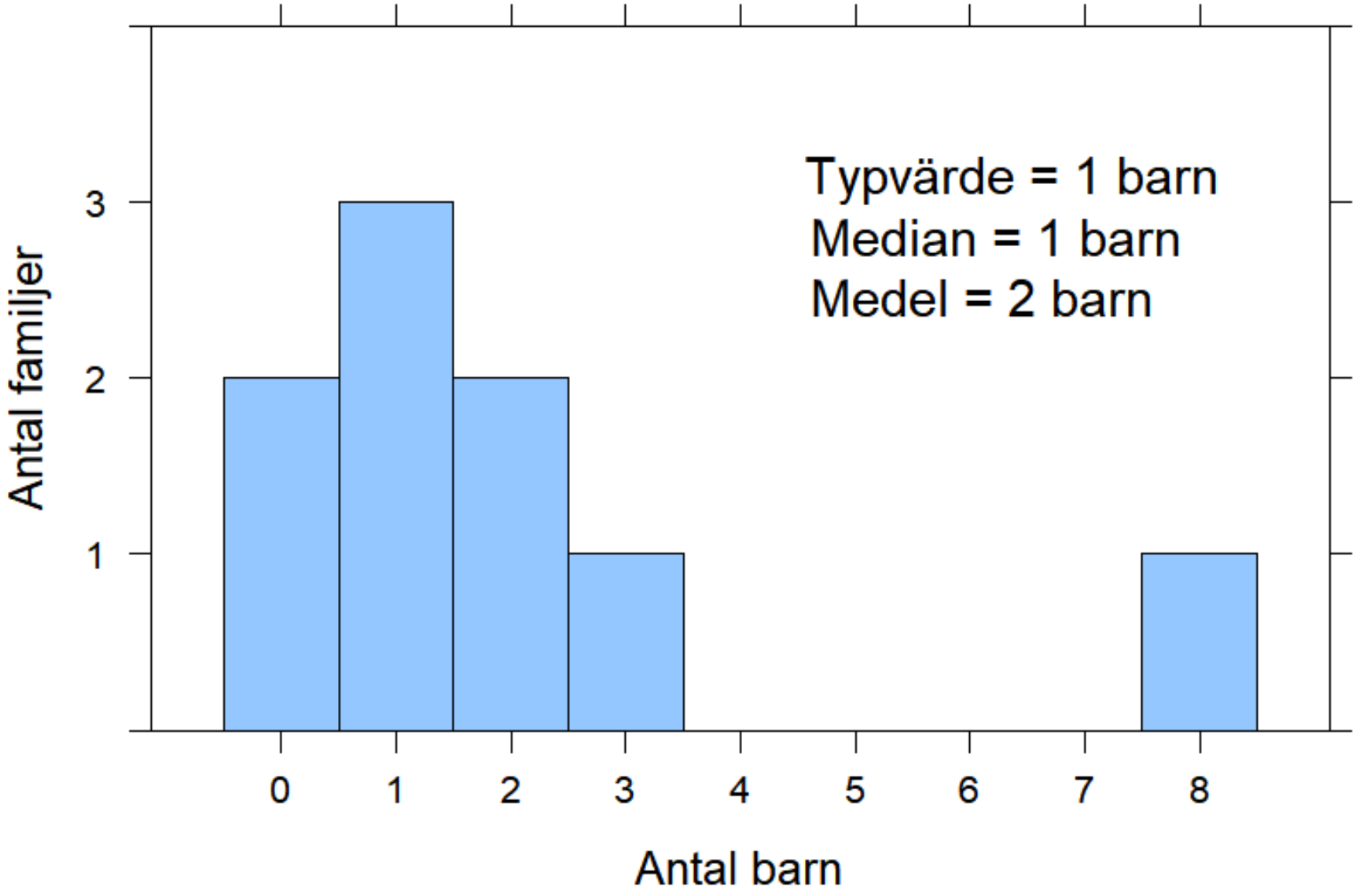
  Median $= 1$

- **Medelvärde** ("mean") - summera alla värden och dividera med n

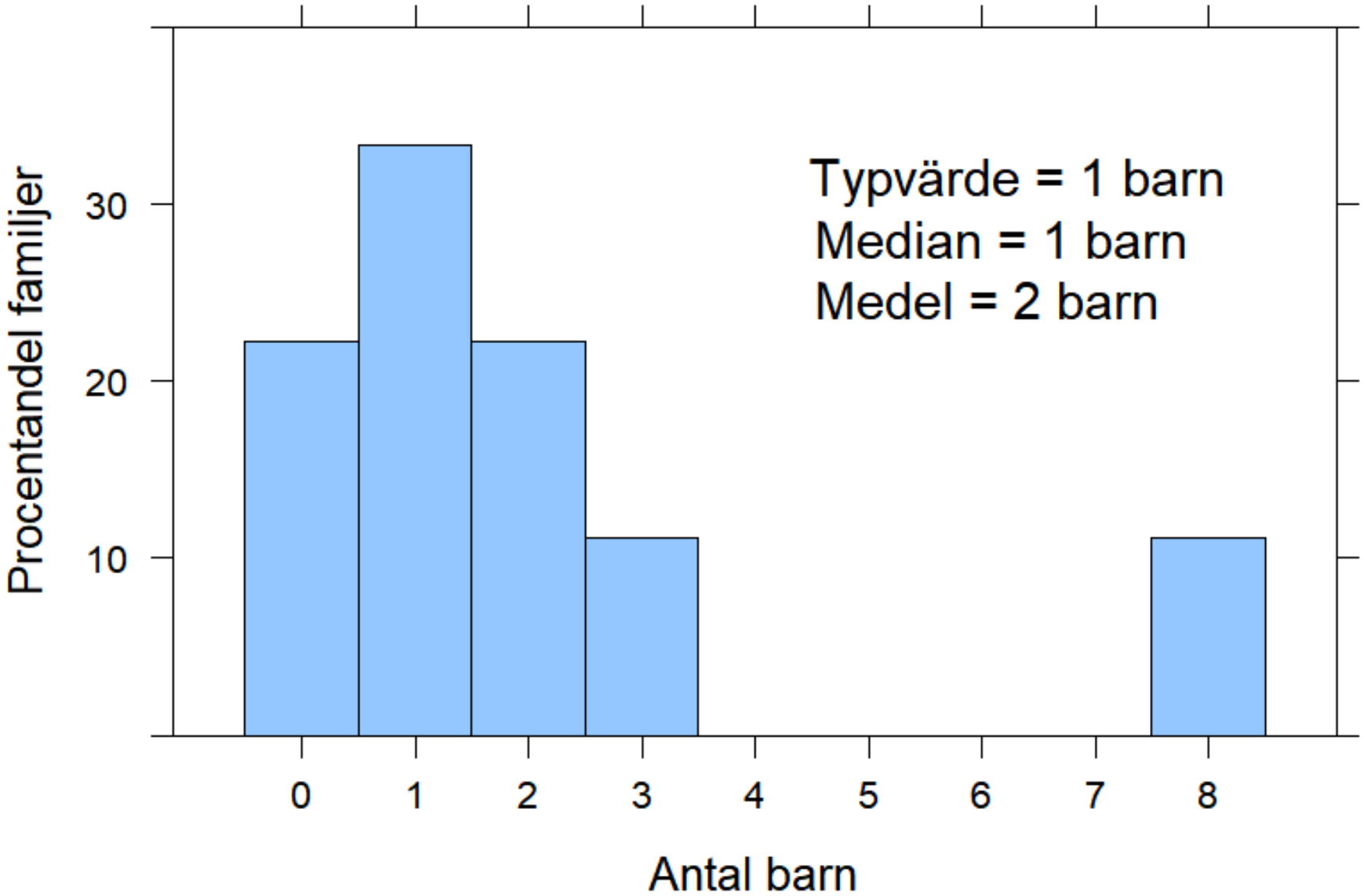
  $\frac{1}{9}\{0 + 3 + 1 + 8 + 1 + 2 + 2 + 0 + 1\}$

  Medel $= \frac{18}{9} = 2$

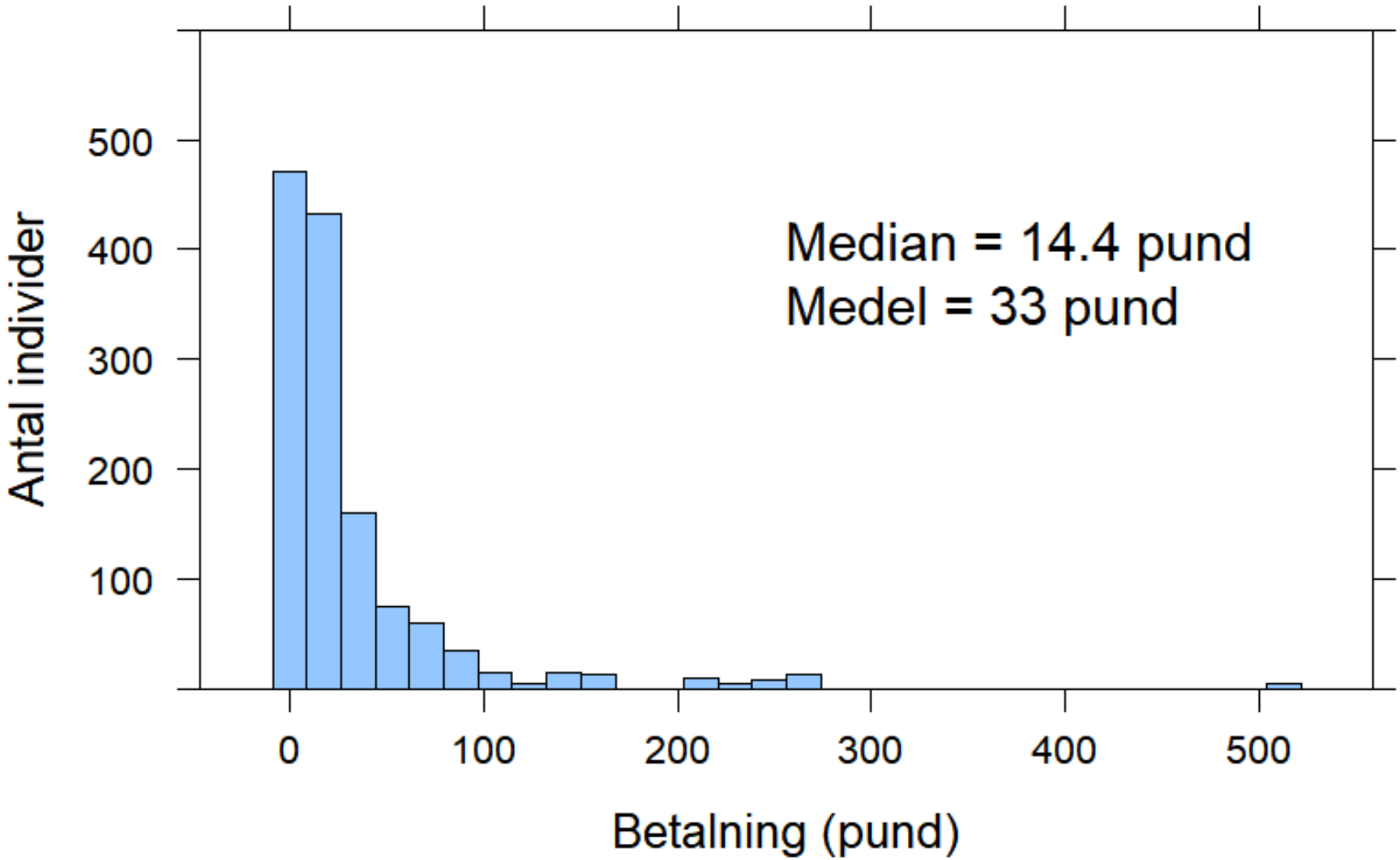
Fördelningen av antal barn i nio familjer



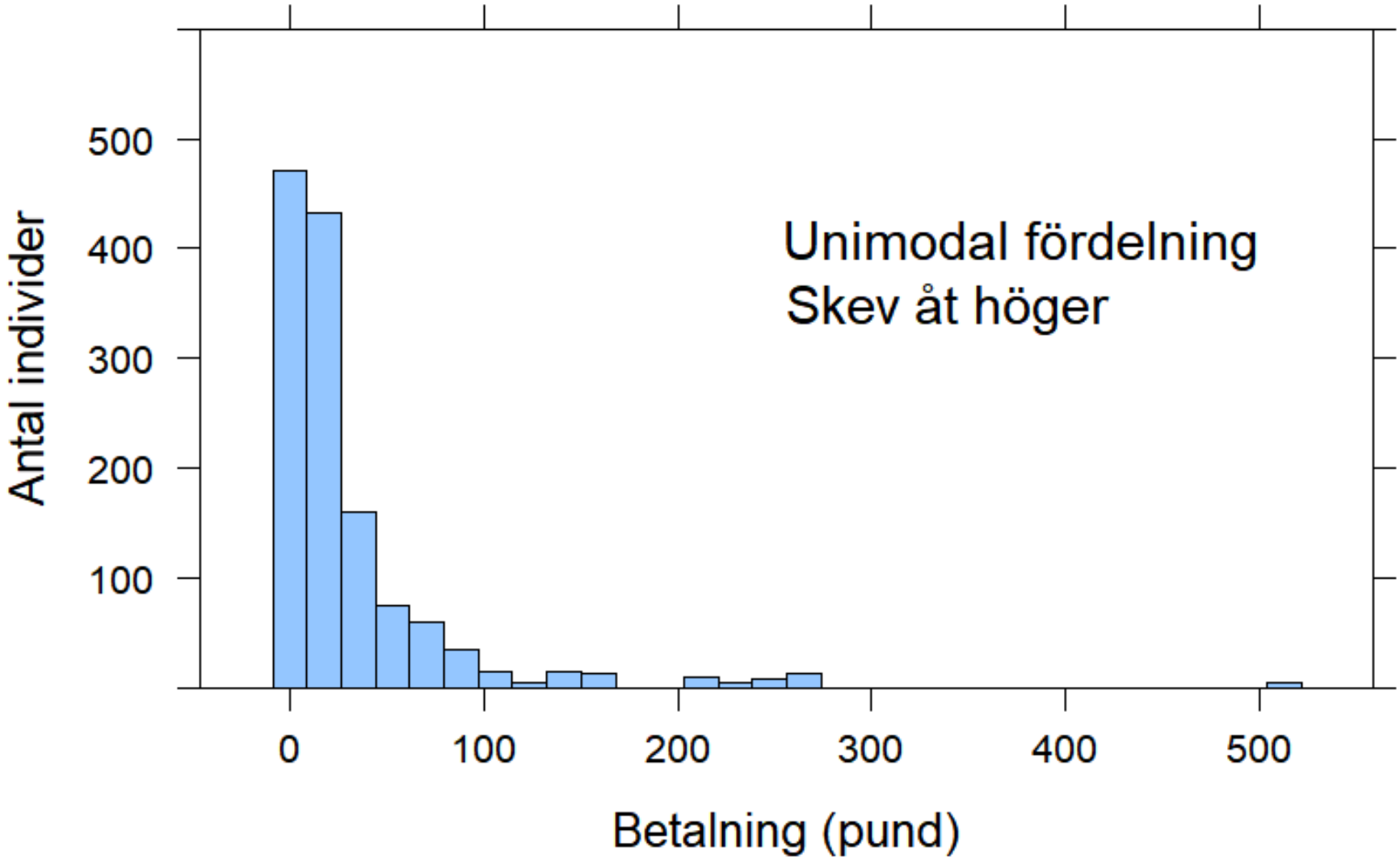
Fördelningen av antal barn i nio familjer



Fördelning av kostnaden för Titanicresan (för 1318 av 2208 passagerare)



Fördelning av kostnaden för Titanicresan (för 1318 av 2208 passagerare)



Median och medelvärde - fortsättning

- **Medianen** - mittersta värdet, efter att observationerna har sorterats
- Om vi har ett jämnt antal observationer, ta medelvärdet av de två värdena i mitten
- **Ex:** Antag att vi nu har data på $n = 8$ familjer
- Medianen av talen $\{0, 0, 1, \mathbf{1}, \mathbf{2}, 2, 3, 8\}$ är 1.5

- **Medelvärdet** av n värden $x_1, x_2, \ldots, x_n$ skrivs som $\bar{x}$ och ges av:

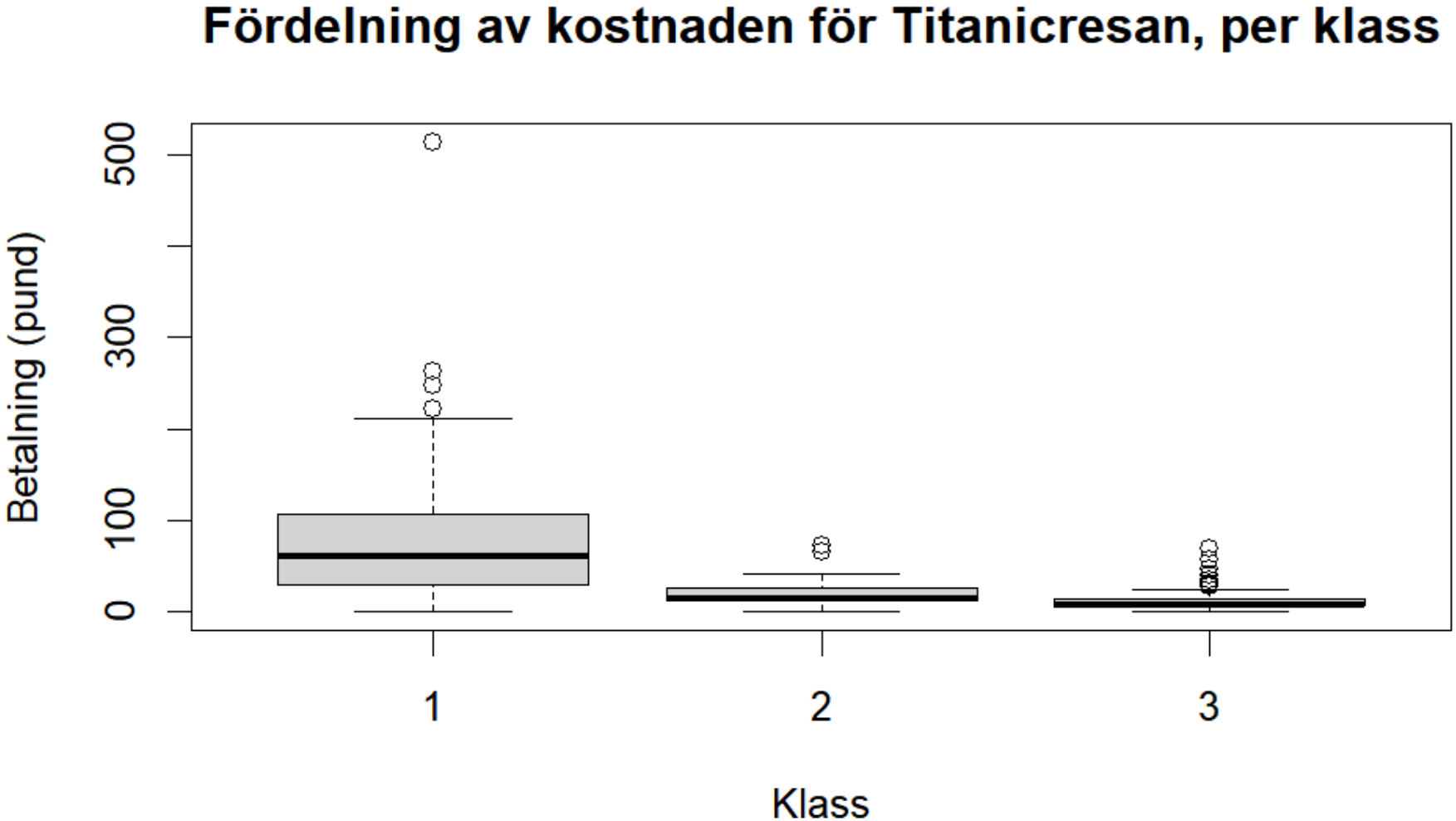
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Vilket/vilka centralmått vi använder beror på syfte (labbuppgift)

- Om du vill repetera algebra, summasymbolen, typvärde / median / medel, etc., se länk på kursens GitHub-sida

Variationsbredd och kvartiler

- Antal barn i åtta familjer: $\{0, 0, 1, 1, 2, 2, 3, 8\}$
- **Variationsbredd** ("Range") - största minus minsta värde ($= 8\text{-}0 = 8$)
- **Kvartiler** ("Quartiles") - dela upp data i fyra lika stora grupper
  - Kvartil 1 - 25% av värdena är mindre, 75% större (0.5)
  - Kvartil 2 - 50% av värdena är mindre, 50% större (1.5) (=median)
  - Kvartil 3 - 75% av värdena är mindre, 25% större (2.5)
- **Interkvartilavstånd** ("Interquartile range, IQR) Tredje minus första kvartilen (2.5-0.5=2)
- Lite olika regler kan användas för exakt hur kvartiler räknas ut
- **Percentiler (fraktiler)** - finare uppdelning
- **Lådagram** (boxplot) - återkommer i labb 2


Fördelning av kostnaden för Titanicresan, per klass
Betalning (pund)
0
100
300
500
1
2
3
Klass

Varians och standardavvikelse

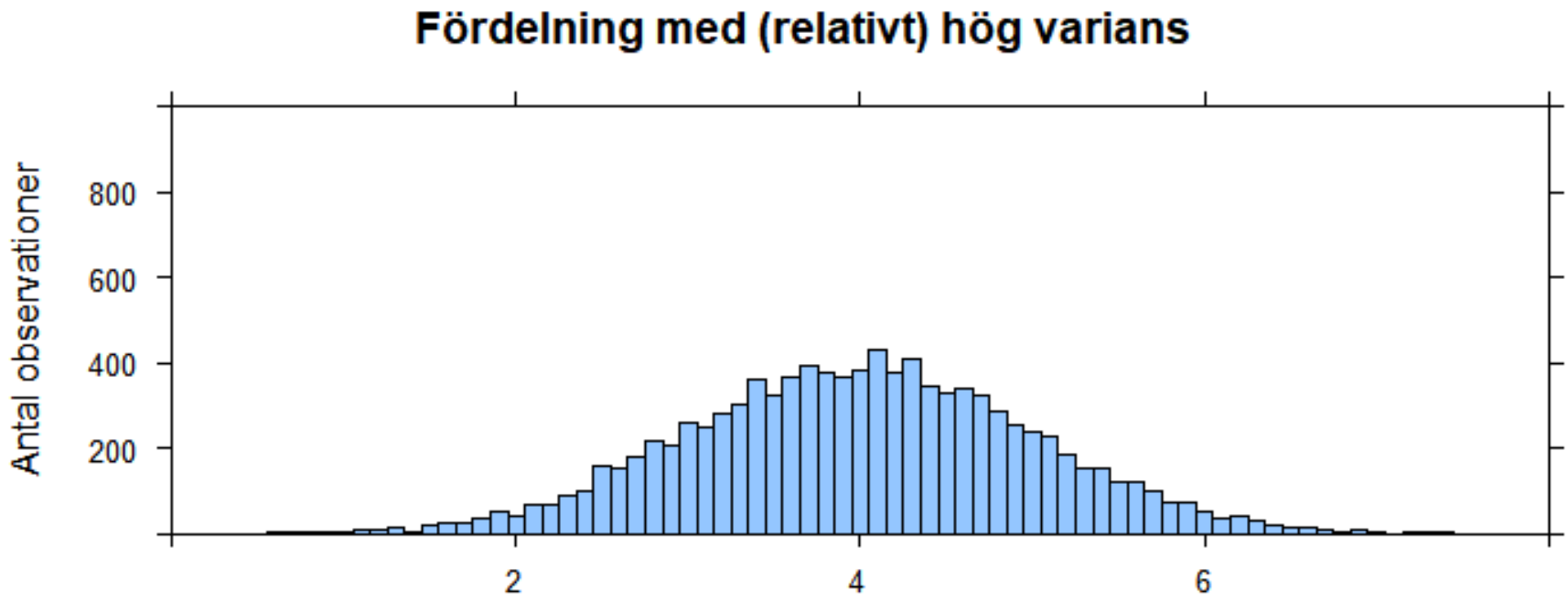
- **Variansen** av n värden $x_1, x_2, \ldots, x_n$ skrivs som s^2 och ges av:

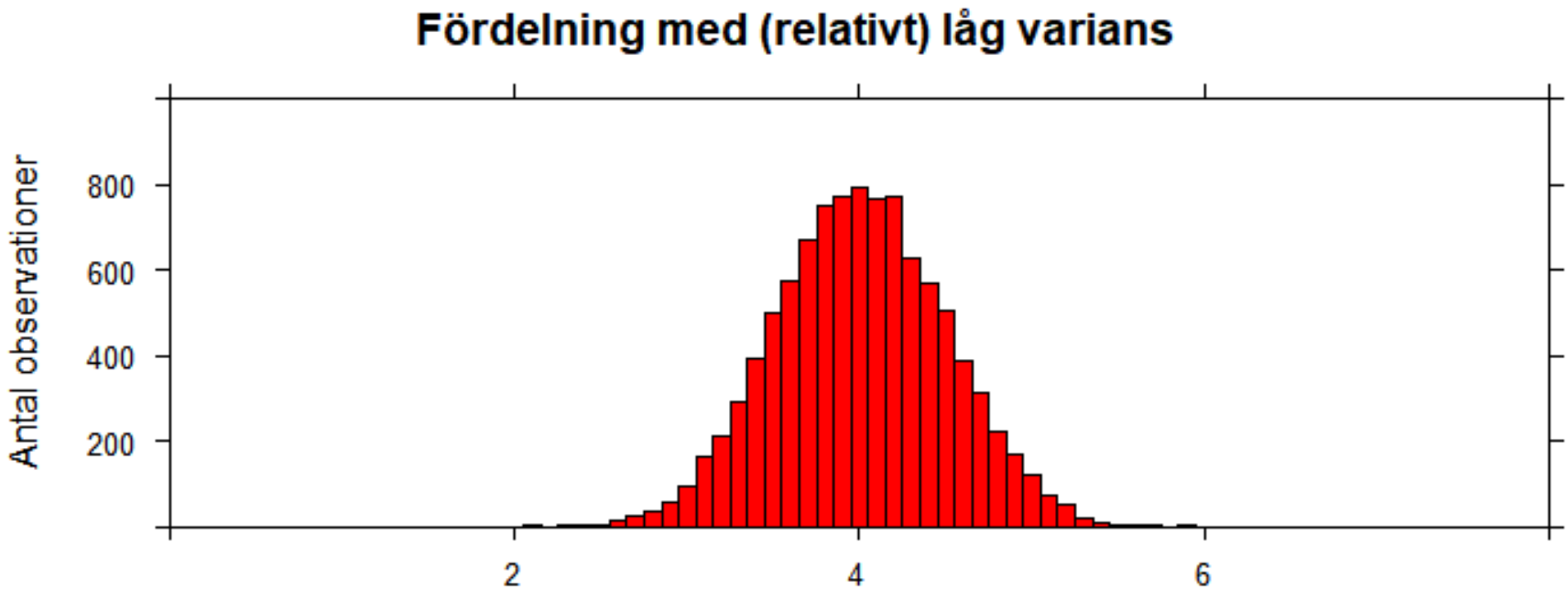
  $$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- **Standardavvikelsen**, kallad s (roten ur variansen), ges av:

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

- Standardavvikelsen är ett mått på hur spridda observationerna är i förhållande till medelvärdet $\bar{x}$.


Fördelning med (relativt) hög varians
Antal observationer
800
600
400
200
2
4
6


Fördelning med (relativt) låg varians
Antal observationer
800
600
400
200
2
4
6

Vidare i kursen

- Ovanstående figur visade två exempel på **normalfördelningen**
- Viktig fördelning som vi återkommer till i föreläsning 4

- Materialet från idag återkommer på datorövningarna 1 och 2

- **Torsdag:** Föreläsning 3 + Datorövning 1 (övningen obligatorisk)
- **Till torsdag:** Testa att installera R (se labb 1)

Denna version av dokumentet: 2025-03-24

Materialet i Statistisk översiktskurs har tagits fram av Ulf Högnäs och Anders Fredriksson, med inspiration och ibland direkt användande av material från andra kurser och personer, bland annat kurserna Statistik och dataanalys 1-3, med material av Michael Carlson, Ellinor Fackle Fornius, Jessica Franzén, Oskar Gustafsson, Oscar Oelrich, Mona Sfaxi, Karl Sigfrid, Mattias Villani, med flera.