# Lecture 2

karl.sigfrid@stat.su.se

# Kursens syfte

Det här är en praktiskt användbar kurs som lär ut hur du

- utvinner insikter ur datamaterial

# Kursens syfte

Det här är en praktiskt användbar kurs som lär ut hur du

- utvinner insikter ur datamaterial
- kommunicerar insikter på ett begripligt sätt

# Kursens syfte

Det här är en praktiskt användbar kurs som lär ut hur du

- utvinner insikter ur datamaterial
- kommunicerar insikter på ett begripligt sätt
- identifierar samband

# Kursens syfte

Det här är en praktiskt användbar kurs som lär ut hur du

- utvinner insikter ur datamaterial
- kommunicerar insikter på ett begripligt sätt
- identifierar samband
- bygger enkla statistiska modeller

# Var används det som kursen lär ut

Det du lär dig här används bland annat av

- data scientists
- business analysts
- statistiker
- ekonomer
- forskare
- ... och alla andra som behöver förstå eller förklara de insikter som ett datamaterial ger

# Tips inför kursen

- Läs igenom kapitlet i boken före föreläsningen.

# Tips inför kursen

- Läs igenom kapitlet i boken före föreläsningen.
- Om ett matematiskt uttryck ser svårt ut, börja med att försöka förstå notationen.**Exempel:** För att förstå innebörden av $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$måste du först förstå vad $\bar{x}$ betyder, vad $\sum_{i=1}^{n} x_i$ betyder och vad står $n$ för.

# Tips inför kursen

- Läs igenom kapitlet i boken före föreläsningen.
- Om ett matematiskt uttryck ser svårt ut, börja med att försöka förstå notationen.**Exempel:** För att förstå innebörden av $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$måste du först förstå vad $\bar{x}$ betyder, vad $\sum_{i=1}^{n} x_i$ betyder och vad står $n$ för.
- Om du fastnar och inte hittar svar i boken, fråga!

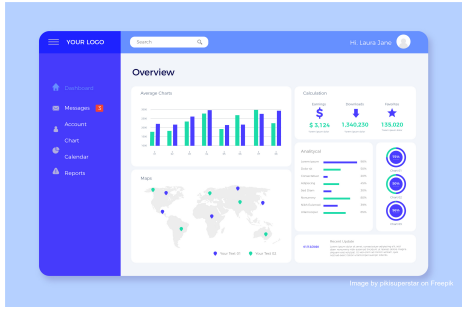
# Tips inför kursen

- Läs igenom kapitlet i boken före föreläsningen.
- Om ett matematiskt uttryck ser svårt ut, börja med att försöka förstå notationen.**Exempel:** För att förstå innebörden av $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$måste du först förstå vad $\bar{x}$ betyder, vad $\sum_{i=1}^{n} x_i$ betyder och vad står $n$ för.
- Om du fastnar och inte hittar svar i boken, fråga!
- Skjut inte upp pluggandet. Börja direkt!

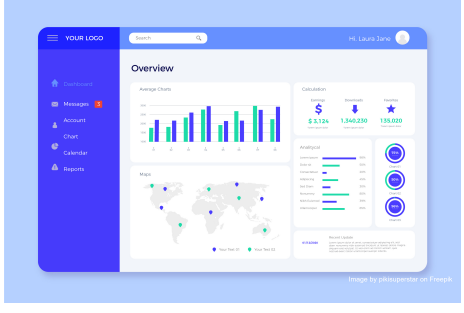
# Två typer av statistik

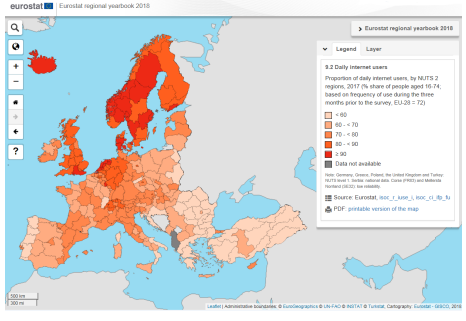
Deskriptiv statistik: Beskriv din data på ett meningsfullt sätt

| Year | Winner | Country of Origin | Age | Team | Total Time (hours) | Avg. Speed (km/h) | Stages | Total Distance Ridden (km) | Starting Riders | Finishing Riders |
|---|---|---|---|---|---|---|---|---|---|---|
| 1903 | Maurice Garin | France | 32 | La Française | 94.55 | 25.7 | 6 | 2428 | 60 | 21 |
| 1904 | Henri Cornet | France | 20 | Cycles JC | 96.10 | 25.3 | 6 | 2428 | 88 | 23 |
| 1905 | Louis Trousselier | France | 24 | Peugeot | 110.45 | 27.1 | 11 | 2994 | 60 | 24 |
| ... | | | | | | | | | | |
| 2011 | Cadel Evans | Australia | 34 | BMC | 86.21 | 39.79 | 21 | 3430 | 198 | 167 |
| 2012 | Bradley Wiggins | Great Britain | 32 | Sky | 87.58 | 39.83 | 20 | 3488 | 198 | 153 |
| 2013 | Christopher Froome | Great Britain | 28 | Sky | 94.55 | 40.55 | 21 | 3404 | 198 | 169 |
| 2014 | Vincenzo Nibali | Italy | 29 | Astana | 89.93 | 40.74 | 21 | 3663.5 | 198 | 164 |
| 2015 | Christopher Froome | Great Britain | 30 | Sky | 84.77 | 39.64 | 21 | 3660.3 | 198 | 160 |
| 2016 | Christopher Froome | Great Britain | 31 | Sky | 89.08 | 39.62 | 21 | 3529 | 198 | 174 |
| 2017 | Christopher Froome | Great Britain | 32 | Sky | 86.34 | 40.997 | 21 | 3540 | 198 | 167 |
| 2018 | Geraint Thomas | Great Britain | 32 | Sky | 83.28 | 40.210 | 21 | 3349 | 176 | 145 |




Inferens: Dra slutsatser om världen utanför






Stockholm University

# Allting börjar med data

Data är allt som vi kan observeras och spara på ett eller annat sätt. Den kan vara strukturerad...

| | mpg | cyl | disp | hp | drat | wt |
|---|---|---|---|---|---|---|
| Mazda RX4 | 21.0 | 6 | 160.0 | 110 | 3.90 | 2.620 |
| Mazda RX4 Wag | 21.0 | 6 | 160.0 | 110 | 3.90 | 2.875 |
| Datsun 710 | 22.8 | 4 | 108.0 | 93 | 3.85 | 2.320 |
| Hornet 4 Drive | 21.4 | 6 | 258.0 | 110 | 3.08 | 3.215 |
| Hornet Sportabout | 18.7 | 8 | 360.0 | 175 | 3.15 | 3.440 |
| Valiant | 18.1 | 6 | 225.0 | 105 | 2.76 | 3.460 |
| Duster 360 | 14.3 | 8 | 360.0 | 245 | 3.21 | 3.570 |
| Merc 240D | 24.4 | 4 | 146.7 | 62 | 3.69 | 3.190 |

... eller ostrukturerad

These days, one of the richest sources of data is the Internet. With a bit of practice, you can learn to find data on almost any subject. Many of the datasets we use in this text were found in this way. The Internet has both advantages and disadvantages as a source of data. Among the advantages are the fact that often you'll be able to find even more current data than those we present. The disadvantage is that references to Internet addresses can "break" as sites evolve, move, and die.

Our solution to these challenges is to offer the best advice we can to help you search for the data, wherever they may be residing. We usually point you to a website. We'll sometimes suggest search terms and offer other guidance.

# Dataset, observationer och variabler

Inom statistikämnet brukar en tabell som denna kallas för ett **dataset**.

| Year | Winner | Country of Origin | Age | Team | Total Time (hours) | Avg. Speed (km/h) | Stages | Total Distance Ridden (km) | Starting Riders | Finishing Riders |
|---|---|---|---|---|---|---|---|---|---|---|
| 1903 | Maurice Garin | France | 32 | La Française | 94.55 | 25.7 | 6 | 2428 | 60 | 21 |
| 1904 | Henri Cornet | France | 20 | Cycles JC | 96.10 | 25.3 | 6 | 2428 | 88 | 23 |
| 1905 | Louis Trousseller | France | 24 | Peugeot | 110.45 | 27.1 | 11 | 2994 | 60 | 24 |
| ... | | | | | .... | | | | | |
| 2011 | Cadel Evans | Australia | 34 | BMC | 86.21 | 39.79 | 21 | 3430 | 198 | 167 |
| 2012 | Bradley Wiggins | Great Britain | 32 | Sky | 87.58 | 39.83 | 20 | 3488 | 198 | 153 |
| 2013 | Christopher Froome | Great Britain | 28 | Sky | 94.55 | 40.55 | 21 | 3404 | 198 | 169 |
| 2014 | Vincenzo Nibali | Italy | 29 | Astana | 89.93 | 40.74 | 21 | 3663.5 | 198 | 164 |
| 2015 | Christopher Froome | Great Britain | 30 | Sky | 84.77 | 39.64 | 21 | 3660.3 | 198 | 160 |
| 2016 | Christopher Froome | Great Britain | 31 | Sky | 89.08 | 39.62 | 21 | 3529 | 198 | 174 |
| 2017 | Christopher Froome | Great Britain | 32 | Sky | 86.34 | 40.997 | 21 | 3540 | 198 | 167 |
| 2018 | Geraint Thomas | Great Britain | 32 | Sky | 83.28 | 40.210 | 21 | 3349 | 176 | 145 |

# Dataset, observationer och variabler

Inom statistikämnet brukar en tabell som denna kallas för ett **dataset**.

| Year | Winner | Country of Origin | Age | Team | Total Time (hours) | Avg. Speed (km/h) | Stages | Total Distance Ridden (km) | Starting Riders | Finishing Riders |
|---|---|---|---|---|---|---|---|---|---|---|
| 1903 | Maurice Garin | France | 32 | La Française | 94.55 | 25.7 | 6 | 2428 | 60 | 21 |
| 1904 | Henri Cornet | France | 20 | Cycles JC | 96.10 | 25.3 | 6 | 2428 | 88 | 23 |
| 1905 | Louis Trousseller | France | 24 | Peugeot | 110.45 | 27.1 | 11 | 2994 | 60 | 24 |
| ... | | | | | .... | | | | | |
| 2011 | Cadel Evans | Australia | 34 | BMC | 86.21 | 39.79 | 21 | 3430 | 198 | 167 |
| 2012 | Bradley Wiggins | Great Britain | 32 | Sky | 87.58 | 39.83 | 20 | 3488 | 198 | 153 |
| 2013 | Christopher Froome | Great Britain | 28 | Sky | 94.55 | 40.55 | 21 | 3404 | 198 | 169 |
| 2014 | Vincenzo Nibali | Italy | 29 | Astana | 89.93 | 40.74 | 21 | 3663.5 | 198 | 164 |
| 2015 | Christopher Froome | Great Britain | 30 | Sky | 84.77 | 39.64 | 21 | 3660.3 | 198 | 160 |
| 2016 | Christopher Froome | Great Britain | 31 | Sky | 89.08 | 39.62 | 21 | 3529 | 198 | 174 |
| 2017 | Christopher Froome | Great Britain | 32 | Sky | 86.34 | 40.997 | 21 | 3540 | 198 | 167 |
| 2018 | Geraint Thomas | Great Britain | 32 | Sky | 83.28 | 40.210 | 21 | 3349 | 176 | 145 |

- Varje rad är en observation.

# Dataset, observationer och variabler

Inom statistikämnet brukar en tabell som denna kallas för ett **dataset**.

| Year | Winner | Country of Origin | Age | Team | Total Time (hours) | Avg. Speed (km/h) | Stages | Total Distance Ridden (km) | Starting Riders | Finishing Riders |
|---|---|---|---|---|---|---|---|---|---|---|
| 1903 | Maurice Garin | France | 32 | La Française | 94.55 | 25.7 | 6 | 2428 | 60 | 21 |
| 1904 | Henri Cornet | France | 20 | Cycles JC | 96.10 | 25.3 | 6 | 2428 | 88 | 23 |
| 1905 | Louis Trousseller | France | 24 | Peugeot | 110.45 | 27.1 | 11 | 2994 | 60 | 24 |
| ... | | | | | .... | | | | | |
| 2011 | Cadel Evans | Australia | 34 | BMC | 86.21 | 39.79 | 21 | 3430 | 198 | 167 |
| 2012 | Bradley Wiggins | Great Britain | 32 | Sky | 87.58 | 39.83 | 20 | 3488 | 198 | 153 |
| 2013 | Christopher Froome | Great Britain | 28 | Sky | 94.55 | 40.55 | 21 | 3404 | 198 | 169 |
| 2014 | Vincenzo Nibali | Italy | 29 | Astana | 89.93 | 40.74 | 21 | 3663.5 | 198 | 164 |
| 2015 | Christopher Froome | Great Britain | 30 | Sky | 84.77 | 39.64 | 21 | 3660.3 | 198 | 160 |
| 2016 | Christopher Froome | Great Britain | 31 | Sky | 89.08 | 39.62 | 21 | 3529 | 198 | 174 |
| 2017 | Christopher Froome | Great Britain | 32 | Sky | 86.34 | 40.997 | 21 | 3540 | 198 | 167 |
| 2018 | Geraint Thomas | Great Britain | 32 | Sky | 83.28 | 40.210 | 21 | 3349 | 176 | 145 |

- Varje rad är en observation.
- Varje kolumn är en variabel.

# Glöm inte att fråga varifrån datamaterialet kommer

- Vi är också intresserade av vad som inom statistikämnet brukar kallas **metadata**. Metadata är information **om** vårt datamaterial.
  - *Vem* har samlat in datamaterialet
  - *Hur* är datamaterialet insamlat?
  - Vad betyder variabelnamnen?
  - Hur är variablerna *kodade*?

# Olika typer av variabler

- Vårt dataset innehåller två typer av variabler:
  - **Kategoriska variabler**
  - **Numeriska variabler** variabler.

| Year | Winner | Country of Origin | Age | Team | Total Time (hours) | Avg. Speed (km/h) | Stages | Total Distance Ridden (km) | Starting Riders | Finishing Riders |
|---|---|---|---|---|---|---|---|---|---|---|
| 1903 | Maurice Garin | France | 32 | La Française | 94.55 | 25.7 | 6 | 2428 | 60 | 21 |
| 1904 | Henri Cornet | France | 20 | Cycles JC | 96.10 | 25.3 | 6 | 2428 | 88 | 23 |
| 1905 | Louis Trousseller | France | 24 | Peugeot | 110.45 | 27.1 | 11 | 2994 | 60 | 24 |
| ... | | | | | .... | | | | | |
| 2011 | Cadel Evans | Australia | 34 | BMC | 86.21 | 39.79 | 21 | 3430 | 198 | 167 |
| 2012 | Bradley Wiggins | Great Britain | 32 | Sky | 87.58 | 39.83 | 20 | 3488 | 198 | 153 |
| 2013 | Christopher Froome | Great Britain | 28 | Sky | 94.55 | 40.55 | 21 | 3404 | 198 | 169 |
| 2014 | Vincenzo Nibali | Italy | 29 | Astana | 89.93 | 40.74 | 21 | 3663.5 | 198 | 164 |
| 2015 | Christopher Froome | Great Britain | 30 | Sky | 84.77 | 39.64 | 21 | 3660.3 | 198 | 160 |
| 2016 | Christopher Froome | Great Britain | 31 | Sky | 89.08 | 39.62 | 21 | 3529 | 198 | 174 |
| 2017 | Christopher Froome | Great Britain | 32 | Sky | 86.34 | 40.997 | 21 | 3540 | 198 | 167 |
| 2018 | Geraint Thomas | Great Britain | 32 | Sky | 83.28 | 40.210 | 21 | 3349 | 176 | 145 |

# Numeriska variabler

| Year | Winner | Country of Origin | Age | Team | Total Time (hours) | Avg. Speed (km/h) | Stages | Total Distance Ridden (km) | Starting Riders | Finishing Riders |
|---|---|---|---|---|---|---|---|---|---|---|
| 1903 | Maurice Garin | France | 32 | La Française | 94.55 | 25.7 | 6 | 2428 | 60 | 21 |
| 1904 | Henri Cornet | France | 20 | Cycles JC | 96.10 | 25.3 | 6 | 2428 | 88 | 23 |
| 1905 | Louis Trousseller | France | 24 | Peugeot | 110.45 | 27.1 | 11 | 2994 | 60 | 24 |
| ... | | | | | .... | | | | | |
| 2011 | Cadel Evans | Australia | 34 | BMC | 86.21 | 39.79 | 21 | 3430 | 198 | 167 |
| 2012 | Bradley Wiggins | Great Britain | 32 | Sky | 87.58 | 39.83 | 20 | 3488 | 198 | 153 |
| 2013 | Christopher Froome | Great Britain | 28 | Sky | 94.55 | 40.55 | 21 | 3404 | 198 | 169 |
| 2014 | Vincenzo Nibali | Italy | 29 | Astana | 89.93 | 40.74 | 21 | 3663.5 | 198 | 164 |
| 2015 | Christopher Froome | Great Britain | 30 | Sky | 84.77 | 39.64 | 21 | 3660.3 | 198 | 160 |
| 2016 | Christopher Froome | Great Britain | 31 | Sky | 89.08 | 39.62 | 21 | 3529 | 198 | 174 |
| 2017 | Christopher Froome | Great Britain | 32 | Sky | 86.34 | 40.997 | 21 | 3540 | 198 | 167 |
| 2018 | Geraint Thomas | Great Britain | 32 | Sky | 83.28 | 40.210 | 21 | 3349 | 176 | 145 |

- Har en enhet (meter, kg, kronor, grader celcius, ...)
- Har storlekar som kan jämföras (2 kg $>$ 1.5 kg)

# Kategoriska variabler

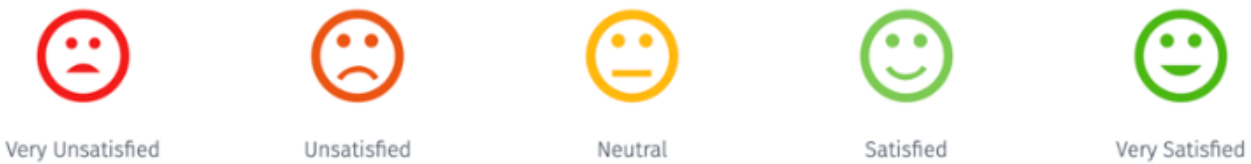
| Year | Winner | Country of Origin | Age | Team | Total Time (hours) | Avg. Speed (km/h) | Stages | Total Distance Ridden (km) | Starting Riders | Finishing Riders |
|---|---|---|---|---|---|---|---|---|---|---|
| 1903 | Maurice Garin | France | 32 | La Française | 94.55 | 25.7 | 6 | 2428 | 60 | 21 |
| 1904 | Henri Cornet | France | 20 | Cycles JC | 96.10 | 25.3 | 6 | 2428 | 88 | 23 |
| 1905 | Louis Trousseller | France | 24 | Peugeot | 110.45 | 27.1 | 11 | 2994 | 60 | 24 |
| ... | | | | | .... | | | | | |
| 2011 | Cadel Evans | Australia | 34 | BMC | 86.21 | 39.79 | 21 | 3430 | 198 | 167 |
| 2012 | Bradley Wiggins | Great Britain | 32 | Sky | 87.58 | 39.83 | 20 | 3488 | 198 | 153 |
| 2013 | Christopher Froome | Great Britain | 28 | Sky | 94.55 | 40.55 | 21 | 3404 | 198 | 169 |
| 2014 | Vincenzo Nibali | Italy | 29 | Astana | 89.93 | 40.74 | 21 | 3663.5 | 198 | 164 |
| 2015 | Christopher Froome | Great Britain | 30 | Sky | 84.77 | 39.64 | 21 | 3660.3 | 198 | 160 |
| 2016 | Christopher Froome | Great Britain | 31 | Sky | 89.08 | 39.62 | 21 | 3529 | 198 | 174 |
| 2017 | Christopher Froome | Great Britain | 32 | Sky | 86.34 | 40.997 | 21 | 3540 | 198 | 167 |
| 2018 | Geraint Thomas | Great Britain | 32 | Sky | 83.28 | 40.210 | 21 | 3349 | 176 | 145 |

- Kan användas för att gruppera observationerna
- Ofta i form av text, men kan vara i form av tal
- En numerisk variabel kan göras till en kategorisk variabel

# Andra typer av variabler

- **Ordinala variabler** kan rangordnas (till skillnad från kategoriska variabler), men har ingen enhet (till skillnad från numeriska variabler).

**Exempel:** Hur nöjd på en femgradig skala är du med ett köp?




- **ID-variabler** har ett unikt värde för varje observation. Kan exempelvis vara personnumret i ett dataset med individer eller årtal i ett dataset där varje observation är ett år.

# Kategoriska variabler - hur de kan beskrivas

Ett klassiskt dataset om passagerare och besättning på skeppet Titanic:

| Name | Survived | Age | Adult/Child | Sex | Price (£) | Class |
|---|---|---|---|---|---|---|
| ABBING, Mr Anthony | Dead | 42 | Adult | Male | 7.55 | 3 |
| ABBOTT, Mr Ernest Owen | Dead | 21 | Adult | Male | 0 | Crew |
| ABBOTT, Mr Eugene Joseph | Dead | 14 | Child | Male | 20.25 | 3 |
| ABBOTT, Mr Rossmore Edward | Dead | 16 | Adult | Male | 20.25 | 3 |
| ABBOTT, Mrs Rhoda Mary "Rosa" | Alive | 39 | Adult | Female | 20.25 | 3 |
| ABELSETH, Miss Karen Marie | Alive | 16 | Adult | Female | 7.65 | 3 |
| ABELSETH, Mr Olaus Jörgensen | Alive | 25 | Adult | Male | 7.65 | 3 |
| ABELSON, Mr Samuel | Dead | 30 | Adult | Male | 24 | 2 |
| ABELSON, Mrs Hannah | Alive | 28 | Adult | Female | 24 | 2 |
| ABRAHAMSSON, Mr Abraham August Johannes | Alive | 20 | Adult | Male | 7.93 | 3 |
| ABRAHIM, Mrs Mary Sophie Halaut | Alive | 18 | Adult | Female | 7.23 | 3 |

Det är svårt att få en bra **överblick** genom att läsa en tabell som den ovan. Vi vet det fanns **2208** personer ombord när skeppet sjönk, men hur kan vi exempelvis få en bra bild av antalet passagerare i varje klass?

# Kategoriska variabler - frekvenstabeller

| Class | Count |
|---|---|
| First | 324 |
| Second | 285 |
| Third | 710 |
| Crew | 889 |

**Table 2.2**
A frequency table of the *Titanic* passengers.

| Class | Percentage (%) |
|---|---|
| First | 14.67 |
| Second | 12.91 |
| Third | 32.16 |
| Crew | 40.26 |

**Table 2.3**
A relative frequency table for the same data.

- En **frekvenstabell** redovisar antalet observationer i varje kategori.
- En **relativ frekvenstabell** visar andel istället för antal.
- Summan i den relativa frekvenstabellen ska bli 100%.$(14.67\% + 12.91\% + 32.16\% + 40.26\% = 100\%)$

# Kategoriska variabler - frekvenstabeller

| Class | Count |
|---|---|
| First | 324 |
| Second | 285 |
| Third | 710 |
| Crew | 889 |

Table 2.2
A frequency table of the *Titanic* passengers.

| Class | Percentage (%) |
|---|---|
| First | 14.67 |
| Second | 12.91 |
| Third | 32.16 |
| Crew | 40.26 |

Table 2.3
A relative frequency table for the same data.

Andelen i procent som tillhör grupp $a$ räknas ut med formeln

$$p_a = \frac{n_a}{n} \cdot 100\%.$$

**Notation:** $p_a$ är andelen i procent som tillhör grupp $a$. $n_a$ är antalet observationer som tillhör grupp $a$, och $n$ är det totala antalet observationer.

**Exempel:** Andelen som tillhörde besättningen var

$$\frac{889}{2208} \cdot 100\% = 40.26\%$$

# Kategoriska variabler - frekvenstabeller

| Class | Count |
|---|---|
| First | 324 |
| Second | 285 |
| Third | 710 |
| Crew | 889 |

**Table 2.2**
A frequency table of the *Titanic* passengers.

| Class | Percentage (%) |
|---|---|
| First | 14.67 |
| Second | 12.91 |
| Third | 32.16 |
| Crew | 40.26 |

**Table 2.3**
A relative frequency table for the same data.

- Den här är vårt första exempel på en **fördelning (distribution)**.
- En fördelning anger, något förenklat, vilka värden en variabel kan ha, och hur ofta varje värde förekommer.
- Fördelning är ett nyckelbegrepp inom statistik.

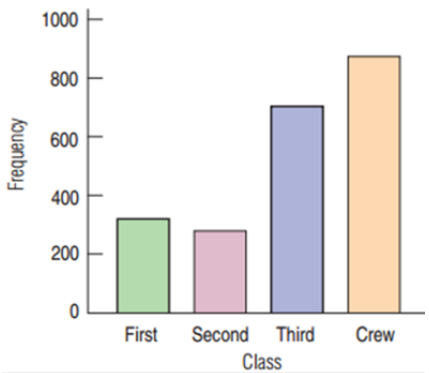
# Kategoriska variabler - grafiska beskrivningar

Vi sammanfatta en variabel mer pedagogiskt med ett diagram:

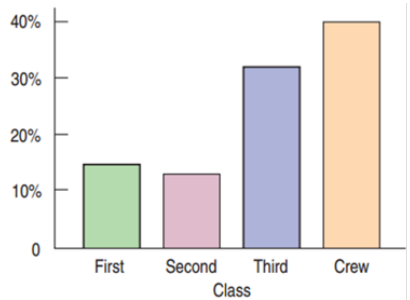
- Att rita diagram har flera fördelar.
  - Vi får en snabbare och tydligare bild av en fördelning.
  - Vi kan se samband som är svåra att se i en tabell.
- För **en kategorisk variabel** kan vi använda
  - **Stapeldiagram (bar plot)**
  - **Pajdiagram (pie chart)**

# Kategoriska variabler - stapeldiagram (bar plot)

- Ett stapeldiagram kan vara **baserat på en frekvenstabell**. Staplarnas höjd anger antalet eller andelen observation som tillhör en viss grupp.
- Ett stapeldiagram kan också vara **baserat på en relativ frekvenstabell**. Staplarnas höjd anger då den andel av observationerna som tillhör en viss grupp.



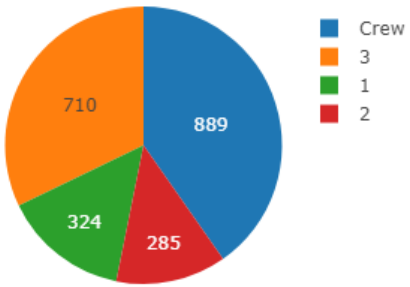

**Frequency Bar Chart**

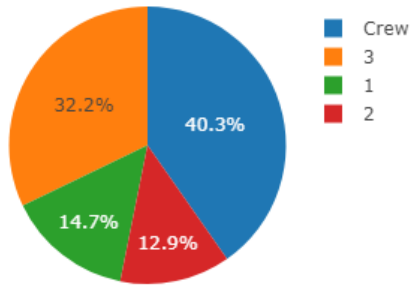


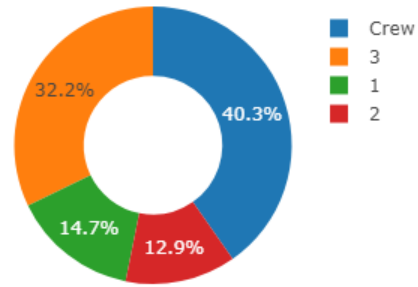

**Relative Frequency Bar Chart**

# Kategoriska variabler - pajdiagram (pie chart)

- Ett pajdiagram
  - fyller samma funktion som ett stapeldiagram.
  - ger en snabb bild av hur stor andel varje grupp utgör.
  - visar tydligt tydligt när andelar är ungefär 1/2 eller 1/4.
- Om det är ett hål i mitten kallas det för ett **munkdiagram (donut chart)**.

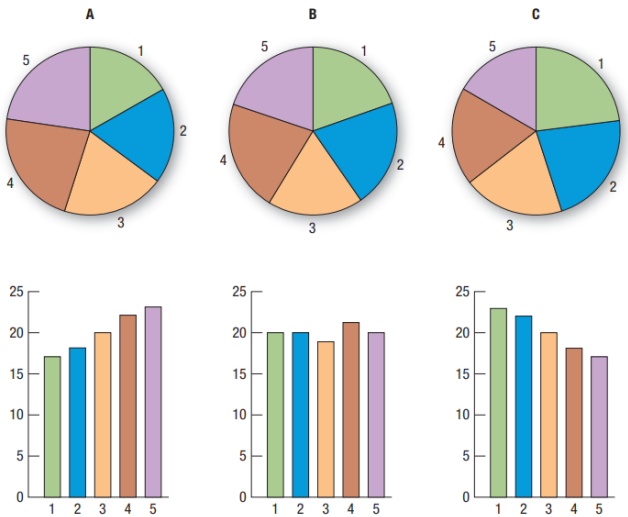


# Kategoriska variabler - stapeldiagram eller pajdiagram?

- Pajdiagram kan vara bättre om publiken har mindre erfarenhet av statistik, medan stapeldiagram brukar föredras av tekniskt kunnig publik.
- I ett stapeldiagram det är lättare att se vilken grupp som är större, särskilt om staplarna står i storleksordning.

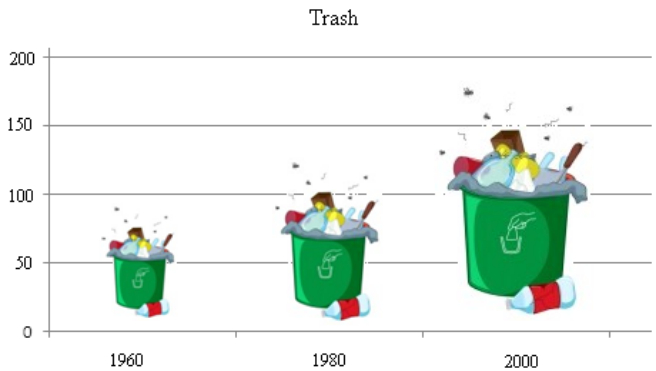


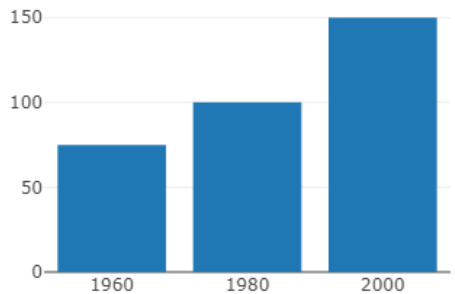
# Kategoriska variabler - areaprincipen

**Fråga:** Jämför den största soptunnan med den minsta? Hur många gånger större skulle du säga att den största soptunnan är?

# Kategoriska variabler - areaprincipen

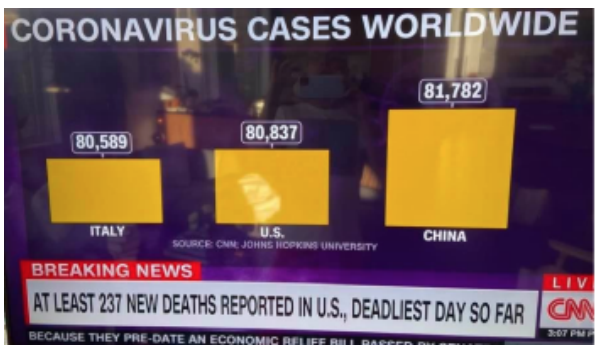
- Vi tenderar att främst se **arean** av en stapel. Därför bör en stapels area vara proportionell till storleken som den representerar. Detta kallas **areaprincipen**.
- Grafen till vänster över hur mängden sopor har ökat är kul, men missvisande. Vi ser på y-axeln att den största soptunnan representerar ungefär **dubbelt** så stor mängd sopor som den minsta, men den är **fyra gånger så stor**. Stapeldiagrammet till höger ger en mer korrekt bild.

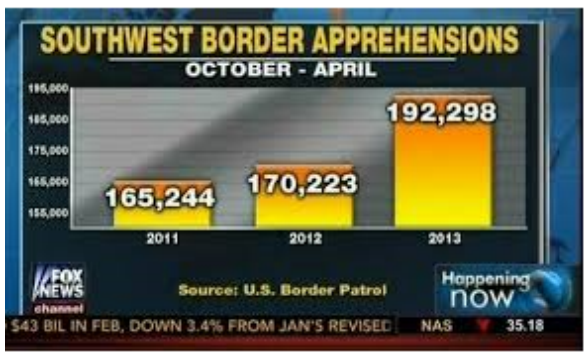


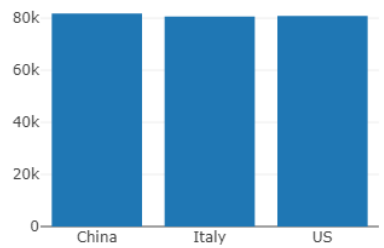



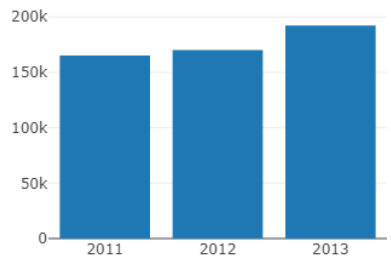
# Kategoriska variabler - areaprincipen

Stapeldiagram bryter ibland mot areaprincipen genom att y-axeln har kapats, dvs y-axeln börjar inte på noll. De nedre diagrammen visar den verkliga relationen mellan staplarna.



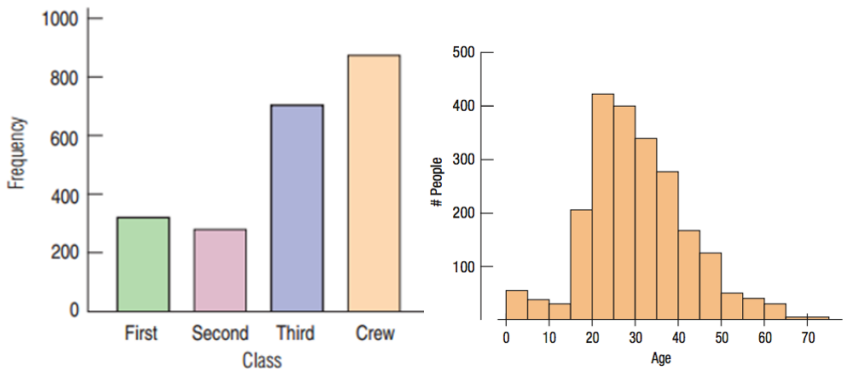






Stockholm University

# Numeriska variabler - hur de kan beskrivas

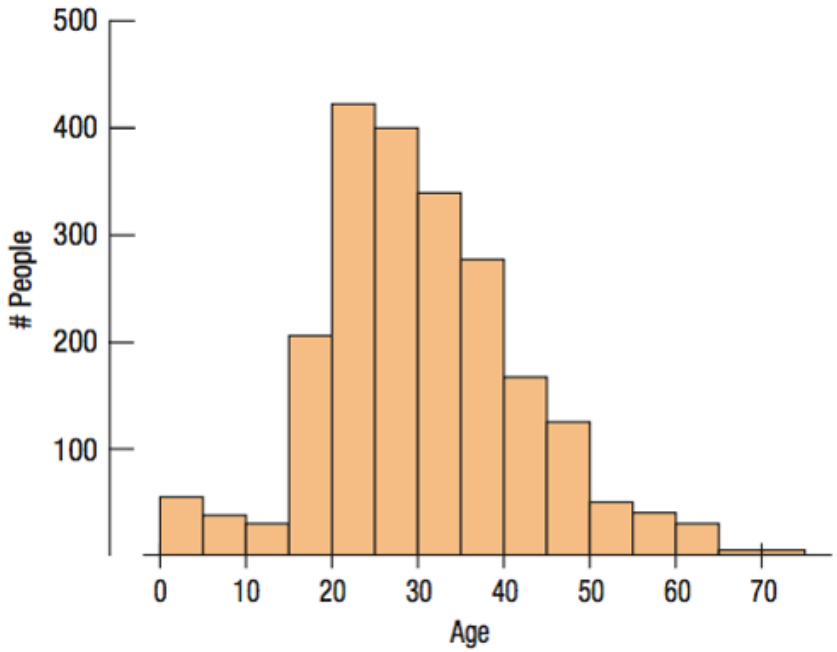
- Stapeldiagram är för kategoriska variabler. För numeriska variabler används **histogram**.
- Histogram ser ut ungefär som stapeldiagram, men istället för kategorier representerar staplarna intervall av numeriska värden.




Till vänster ett stapeldiagram för den kategoriska variabeln *Class*. Till höger ett histogram för den numeriska variabeln *Age*.

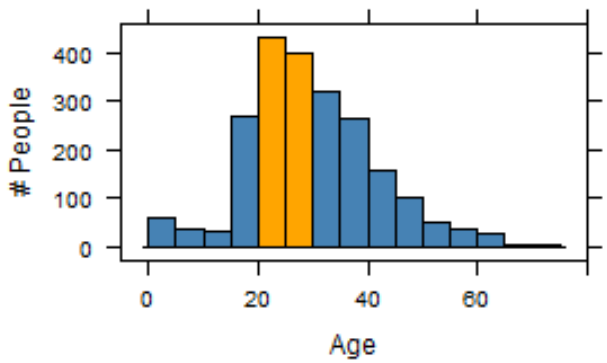
# Numeriska variabler - histogram

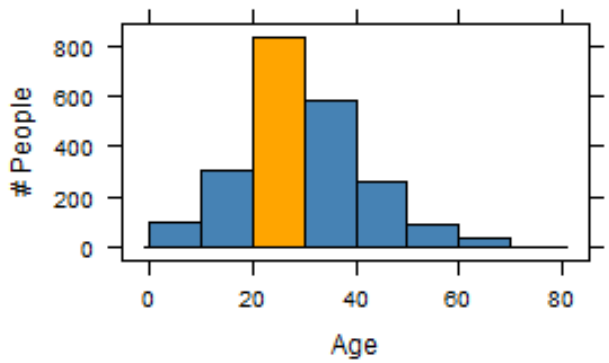
- Varje stapel representerar ett åldersintervall på 5 år.
- Vi ser vi att alla passagerare på Titanic var mellan 0 och 75 år gamla.
- Den fjärde stapeln från vänster visar att drygt 200 passagerare var 15-19 år.



# Numeriska variabler - histogram

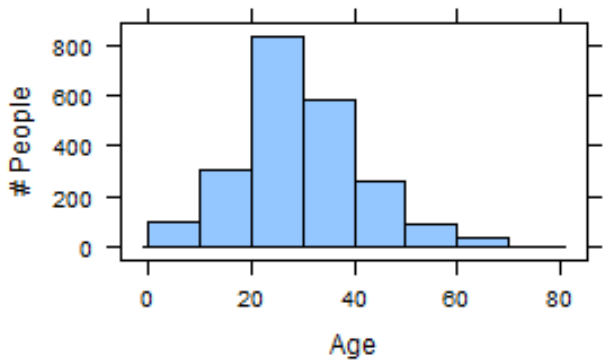
- När du gör ett histogram väljer du själv **bredden** på dina intervall.
- Notera att höjden på de två markerade staplarna i det vänstra histogrammet representerar ungefär 400 personer vardera. I det högra histogrammet är de båda staplarna ihopslagna, och den sammanslagna stapeln representerar då ungefär 800 personer.

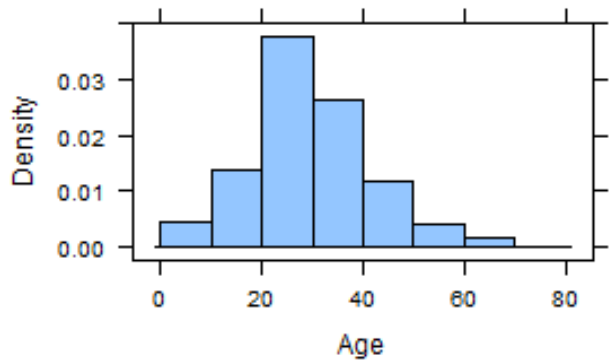


# Numeriska variabler - histogram

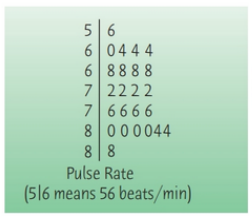
- Det finns också histogram som kallas **täthetshistogram (density histogram)**. I ett sådant histogram prepresenterar **arean** av en stapel den andel av observationerna som ligger inom stapels intervall.
- **Exempel**: Den högsta stapeln i figuren till höger (20-29 år) har en höjd som är ungefär 0.036. Stapelns bredd är 10, så arean är $0.036 * 10 = 0.36$. Andelen personer på Titanic som var i åldern 20-29 år var alltså ungefär 36%.

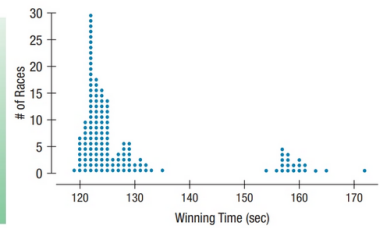


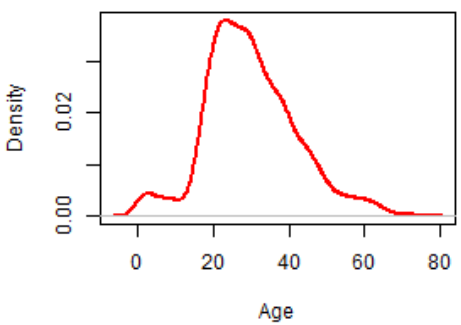



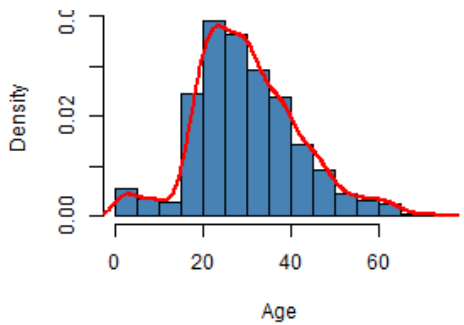
# Numeriska variabler - ytterligare typer av diagram

- Det finns även stam- och bladdiagram (överst till vänster), punktdiagram (överst till höger) och täthetdiagram (underst).
- Täthetsdiagrammet har samma form som ett histogram, men är utjämnat.





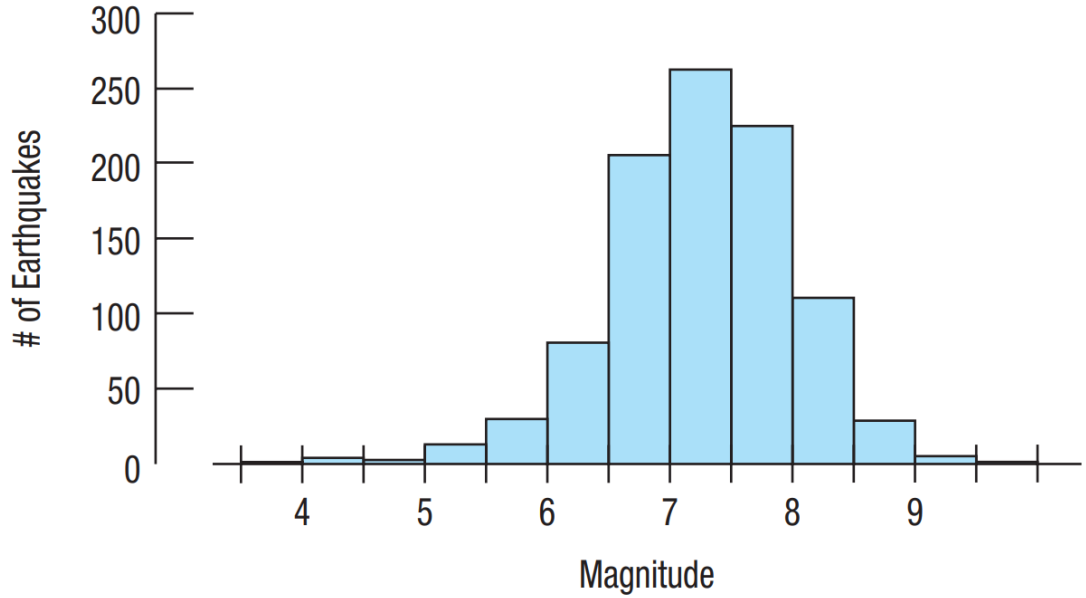

Stockholm University

# Numeriska variabler - att analysera histogram

- Formen på ett histogram kan ge oss intressant information om hur värden på en variabel fördelar sig i ett dataset.
- **Typvärdet (mode)** är det värde av en variabel som har det största antalet observationer. Det representeras av toppen av fördelningskurvan.
- **Symmetrin (symmetry/skewness)** anger om fördelningen är symmetrisk eller sned.
- **Extrema värden (outliers)** är observationer som ligger långt från övriga observationer.

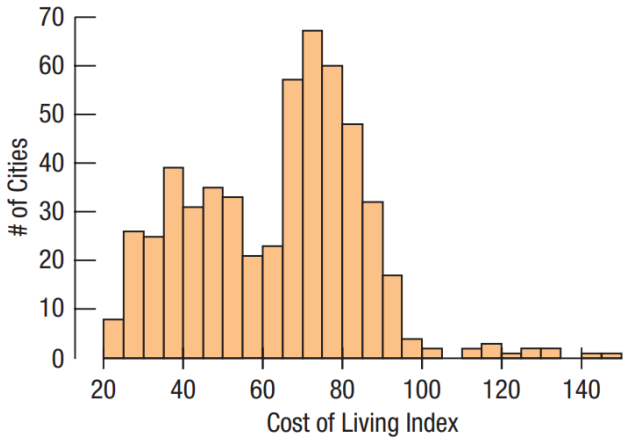
# Numeriska variabler - typvärde

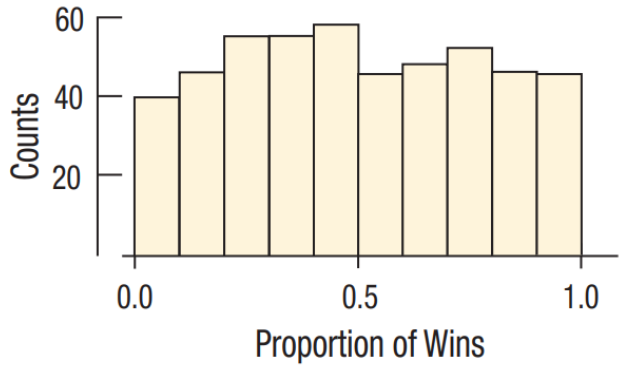
- Om fördelningen av en variabel har en enda topp så hittar vi typvärdet där. En sådan fördelning är **unimodal**. Figuren, som visar magnituden på jordbävningar, har sitt typvärde i närheten av 7.



# Numeriska variabler - typvärde

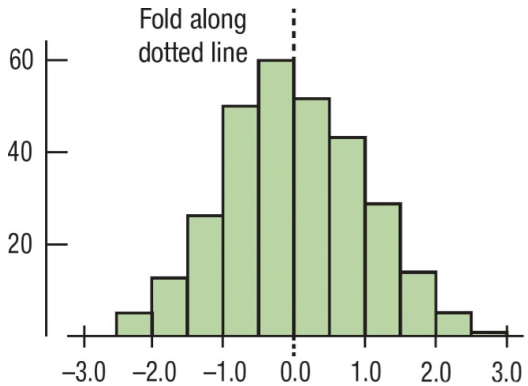
- En fördelning med två toppar är **bimodal** och har den fler toppar är den **multimodal**. Figuren till vänster, som visar ett index för levnadskostnader i olika städer, har en topp vid 40 och en vid 80. Kanske döljer sig två olika grupper av städer i datamaterialet.
- En fördelning som är jämn utan tydliga toppar och dalar, som den till höger, kallas för en **uniform fördelning**.

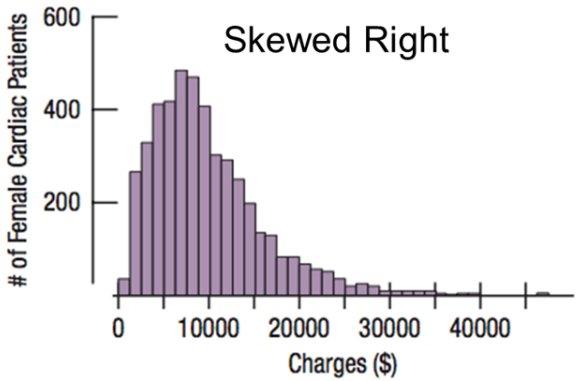


# Numeriska variabler - symmetri

- Det **gröna** histogrammet är **symmetriskt**. Den högra halvan av histogrammet är på ett ungefär en spegelbild av den vänstra.
- **Det lila** histogrammet, som visar hur mycket kvinnliga hjärtpatienter har fakturerats, är **skevt åt höger (right skewed)**. Det kan tolkas som att många patienterna har fakturerats en summa högt över typvärdet, medan få har fakturerats långt under typvärdet.



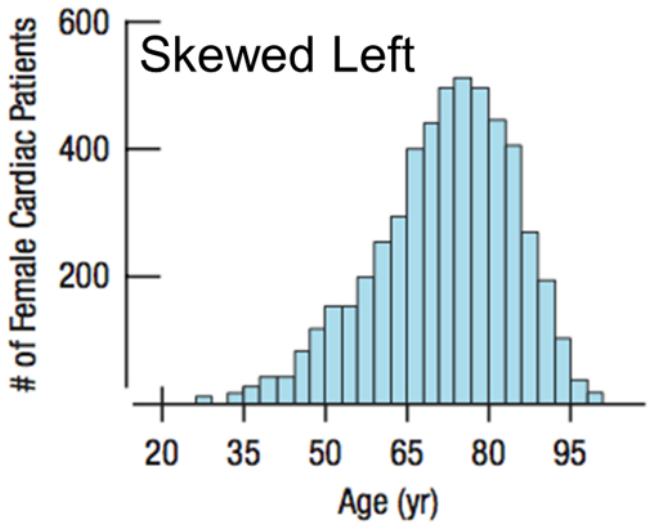






Stockholm University

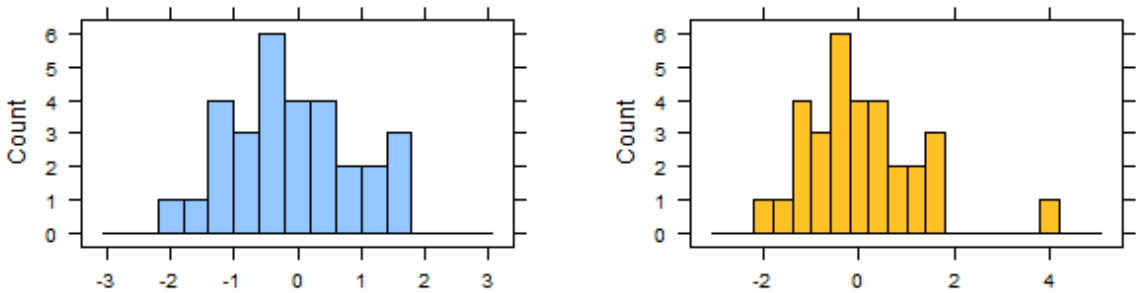
# Numeriska variabler - symmetri

- Det **blå** histrogrammet, som visar åldern hos kvinnliga hjärtpatienter, är **skevt åt vänster (left skewed)**. Det kan tolkas som att många patienter har en ålder långt under typvärdet, men få har en ålder långt över typvärdet.



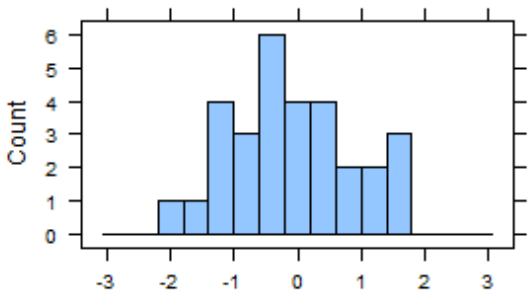
# Numeriska variabler - outliers

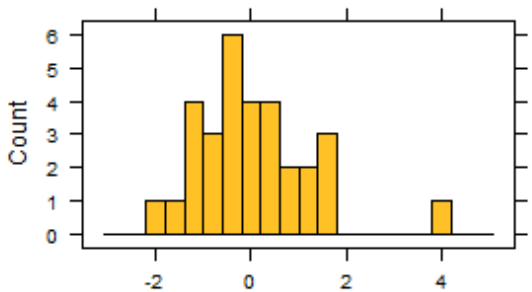
- Extrema värden som avviker från övriga observationer brukar kallas för **outliers**, även på svenska.
- Det blå histogrammet nedan har inga outliers. Alla Observationer ligger samlade. Det gula histogrammet har en outlier till höger om de övriga observationerna.



# Numeriska variabler - outliers

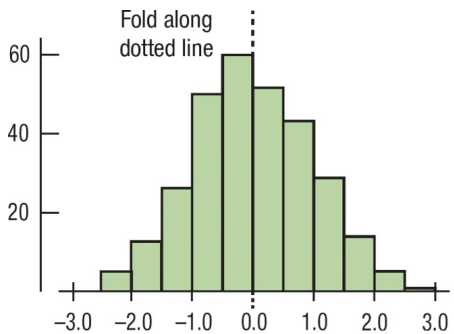
- Outliers kan få stora effekter i en statistisk analys.
- Outliers behöver ofta utredas. De kan finnas där på grund av misstag i datainsamlingen, men de kan också vara korrekta observationer.
- Om outliers tas bort ur datamaterialet måste detta dokumenteras och motiveras.

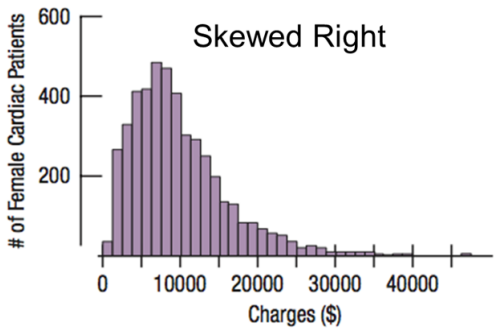


# Numeriska variabler - fördelningens centrum

- Vi vill ofta ha ett mått på det typiska värdet av en variabel. Vi är då ute efter ett värde vid fördelningens centrum.
- Det typiska värdet identifieras enkelt i mitten av en symmetrisk fördelning.
- Om fördelningen är skev är det svårare att säga vad som ska rapporteras som ett typiskt värde.
- Tre alternativa mått som beskriver var fördelningens centrum ligger är **typvärde (mode)**, **medelvärde (mean)**, och **median**.



# Numeriska variabler - medelvärde (mean)

Anta att vi har 7 observationer av en variabel som vi kallar $x$:

$$x_1 = 12, x_2 = 11, x_3 = 9, x_4 = 13, x_5 = 12, x_6 = 10, x_7 = 11$$

Medelvärdet av de här observationerna är

$$\frac{12 + 11 + 9 + 13 + 12 + 10 + 11}{7} = 11.14$$

Mer allmänt kan vi säga att medelvärdet beräknas

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

# Numeriska variabler - medelvärde

Låt oss förklara notationen i uttrycket

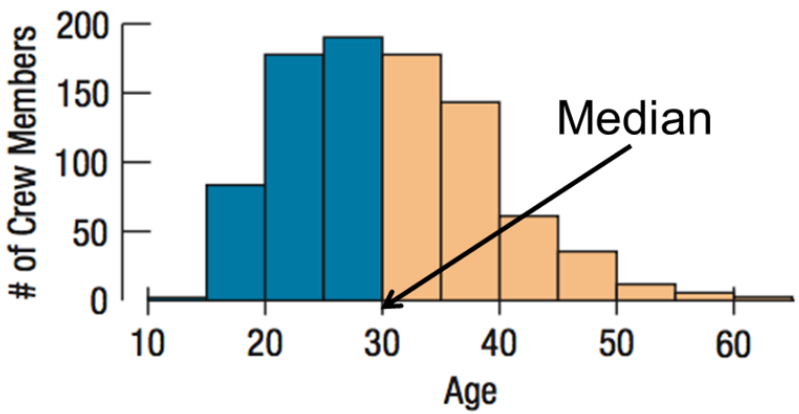
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

- Uttrycket $\bar{x}$ är medelvärdet för variabeln $x$. Ett $x$ med en linje över kan uttalas *"x bar"*. På samma sätt är $\bar{y}$ medelvärdet för variabeln $y$, osv.
- Bokstaven $n$ brukar användas som symbol för antalet observationer i vårt dataset. Om vi, som i exemplet, har 7 observationer är $n = 7$.
- Uttrycket $\sum_{i=1}^{n} x_i$ är summan av alla värden av variabeln $x$, dvs

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + ... + x_n$$

# Numeriska variabler - median

- Medianen är ett värde som är större än ungefär hälften av observationerna och mindre än ungefär hälften av observationerna. Att vi säger *ungefär* beror på att antalet observationer inte alltid är jämnt delbart med 2.
- Figuren visar åldersfördelningen för Titanics besättning. Anta att de **blå** staplarna i figuren representerar lika många personer som staplarna i **beige**. Antalet besättningsmän på Titanic som är under 30 år är då lika stort som antalet över 30 år. Medianåldern är alltså omkring 30 år.



# Numeriska variabler - median

Vi hittar medianen på följande sätt:

1. Sortera observationerna från lägsta till högsta värde.
2. Om antalet observationer är udda, identifiera den mittersta observationen. Värdet på denna är detsamma som medianen. Om antalet observationer är jämnt, identifiera de två observationerna som ligger i mitten. Medelvärdet av dessa två är detsamma som medianen.

# Numeriska variabler - median

**Exempel med udda antal observationer**

Vi har variabeln $x$ med följande 5 värden:

| x | | | | |
|---|---|---|---|---|
| 14.7 | 2.2 | 1.7 | 3.09 | 3.11 |

Vi börjar med att sortera våra värden i storleksordning.

| x | | | | |
|---|---|---|---|---|
| 1.7 | 2.2 | 3.09 | 3.11 | 14.7 |

Värdet i mitten av den sorterade listan är 3.09, så medianen är **3.09**.

# Numeriska variabler - median

**Exempel med jämnt antal observationer**

Vi har variabeln $x$ med följande 6 värden:

| x | | | | | |
|---|---|---|---|---|---|
| 14.7 | 2.2 | 1.7 | 3.09 | 3.11 | 16.3 |

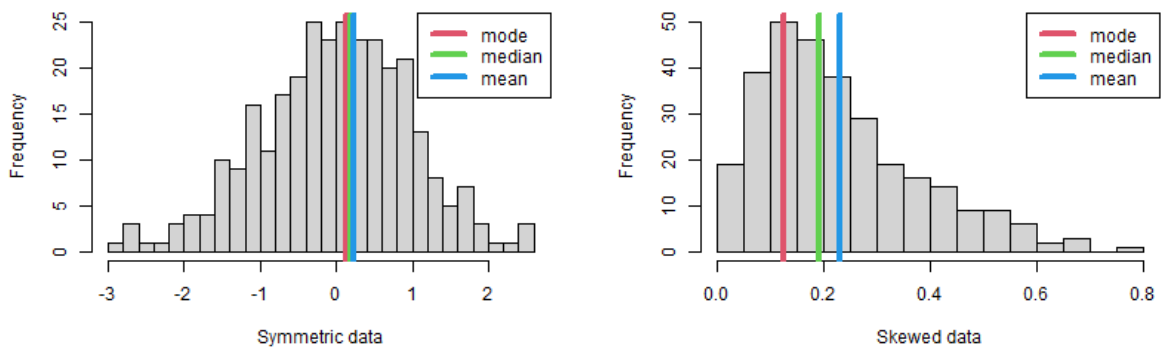
Vi börjar med att sortera våra värden i storleksordning.

| x | | | | | |
|---|---|---|---|---|---|
| 1.7 | 2.2 | 3.09 | 3.11 | 14.7 | 16.3 |

De två värden som ligger i mitten av listan är 3.09 och 3.11. Medelvärdet av dessa värden är $(3.09 - 3.11)/2 = 3.10$. Medianen är alltså **3.10**.

# Numeriska variabler - median, medelvärde eller typvärde

- Om fördelningen är symmetrisk, som i bilden till vänster brukar skillnaden vara liten mellan de olika måtten.
- Om fördelningen är skev, som i bilden till höger, påverkas medelvärdet mer av värden ute i svansarna.
- Outliers kan kan stor påverkan på medelvärdet, men inte på medianen.

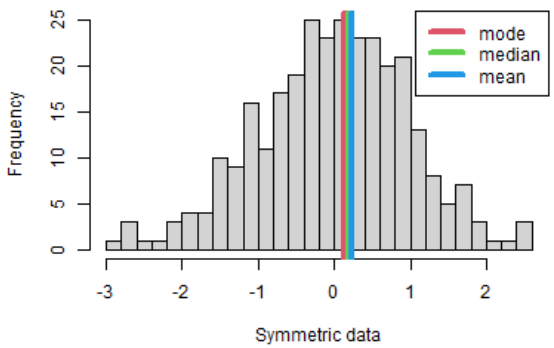


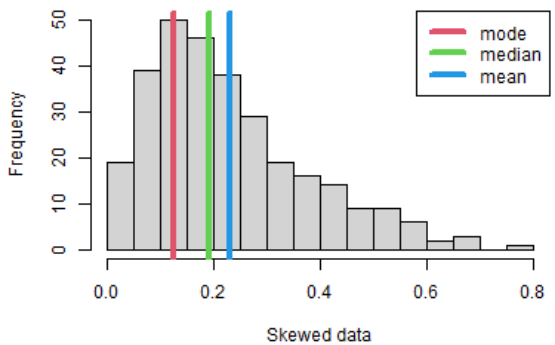



Stockholm University

# Numeriska variabler - median, medelvärde eller typvärde

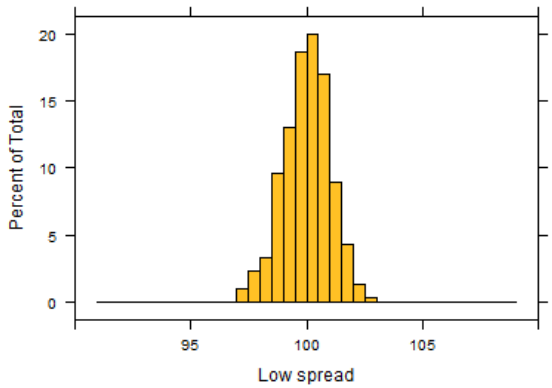
- Vilket värde som bör rapporteras beror på syftet.
- Om du säljer biljetter till en båtresa och vill du veta hur stora intäkterna blir är du förmodligen intresserad av medelvärdet av biljettpriset.
- En köpare som undrar vad en typisk biljett kostar är kanske mer intresserad av medianpriset, som inte påverkas av priset på de allra dyraste biljetterna.

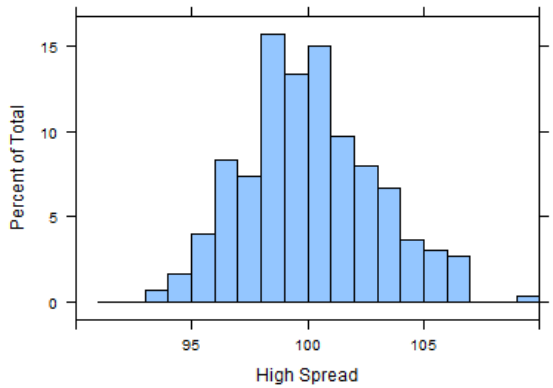


# Numeriska variabler - fördelningens spridning

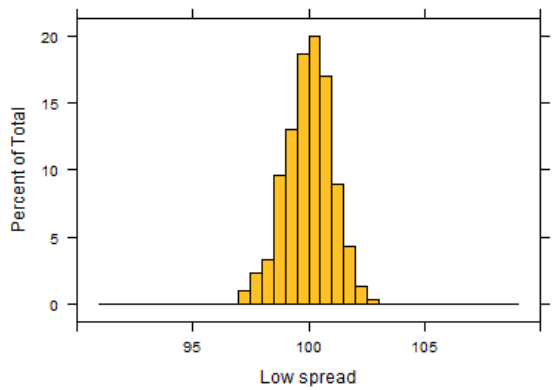
- Den gula och den ljusblå fördelningen har ungefär samma medelvärde, men olika **spridning**.
- Om histogrammen visar inkomstfördelningen i två länder så representerar det gula ett land där lönenivåerna är mer lika, och det ljusblå ett land där inkomstskillnaderna är större.

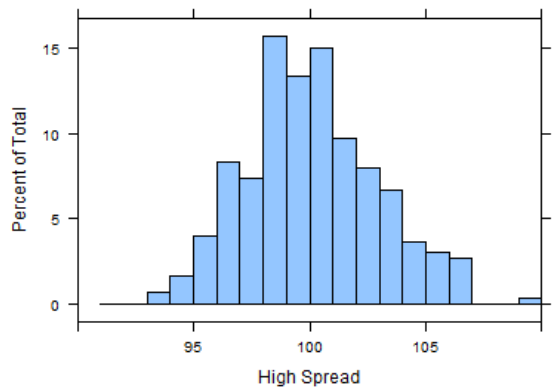


# Numeriska variabler - fördelningens spridning

- Det finns olika mått på hur stor spridningen är:
  - Variationsbredd (range)
  - Standardavvikelse (standard deviation)
  - Kvartilavstånd (interquartile range)



# Numeriska variabler - Variationsbredd

- **Variationsbredden** mäter avståndet mellan den största och den minsta observationen.
- Variationsbredden påverkas kraftigt av outliers.

**Exempel**

Bland Titanics besättningsmän var den äldsta 62 år och den yngsta 14. Variationsbredden för den åldersvariabeln är alltså $62 - 14 = 48$.

# Numeriska variabler - Standardavvikelse

**Standardavvikelsen** anger hur mycket observationerna avviker från medelvärdet.

För att räkna ut standardavvikelsen är det lättast att först räkna ut **variansen**, som vi betecknar $s^2$. Variansen, som är standardavvikelsen i kvadrat, räknas ut med formeln

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1},$$

där $n$ är antalet observationer.

Standardavvikelsen, som vi betecknar $s$, är kvadratroten ur variansen:

$$s = \sqrt{s^2}$$

# Numeriska variabler - Standardavvikelse

Låt oss förklara notationen

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}.$$

- Uttrycket $(y_i - \bar{y})^2$ är den kvadrerade skillnaden mellan observationen $y_i$ och medelvärdet av $y$.
- Uttrycket $\sum_{i=1}^{n}(y_i - \bar{y})^2$ är alltså summan av dessa kvadrerade skillnader.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + ... + (y_n - \bar{y})^2,$$

där $y_n$ är vår sista observation. För att kunna göra uträkningen måste vi först ha räknat ut medelvärdet $\bar{y}$.

# Numeriska variabler - Standardavvikelse, exempel, del1

Anta att vi har nio säckar med jord, och som ett mått på hur mycket vikten skiljer sig åt mellan säckarna vill vi räkna ut standardavvikelsen.

- Variabeln $y$ på första raden i nedanstående tabell anger vikten på varje säck i kg. Medelvikten är $\bar{y} = (23 + 27 + 22 + 11 + 18 + 26 + 19 + 13 + 28)/9 = 20.78$ kg.
- På andra raden i tabellen räknar vi ut $(y_i - \bar{y})^2$ för varje observation. För $y_1$ får vi till exempel $(23 - 20.78)^2 = 4.93$. För $y_2$ får vi $(27 - 20.78)^2 = 38.69$.
- På tredje raden räknar vi ut summan av alla värden från andra raden. $\sum_{i=1}^{9}(y_i - \bar{y})^2 = 4.93 + 38.69 + 1.49 + 95.65 + 7.73 + 27.25 + 3.17 + 60.53 + 52.13 = 291.57$

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $y_7$ | $y_8$ | $y_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 23 | 27 | 22 | 11 | 18 | 26 | 19 | 13 | 28 |
| $(y_i - \bar{y})^2$ | 4.93 | 38.69 | 1.49 | 95.65 | 7.73 | 27.25 | 3.17 | 60.53 | 52.13 |
| $\sum_{i=1}^{9}(y_i - \bar{y})^2$ | | | | | | | | | 291.57 |

# Numeriska variabler - Standardavvikelse, exempel, del2

Vi har räknat ut att $\sum_{i=1}^{9}(y_i - \bar{y})^2 = 291.57$, och vi vet att antalet observationer är $n = 9$. Om vi sätter in våra värden i formeln får vi

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{291.57}{9-1} = 36.446$$

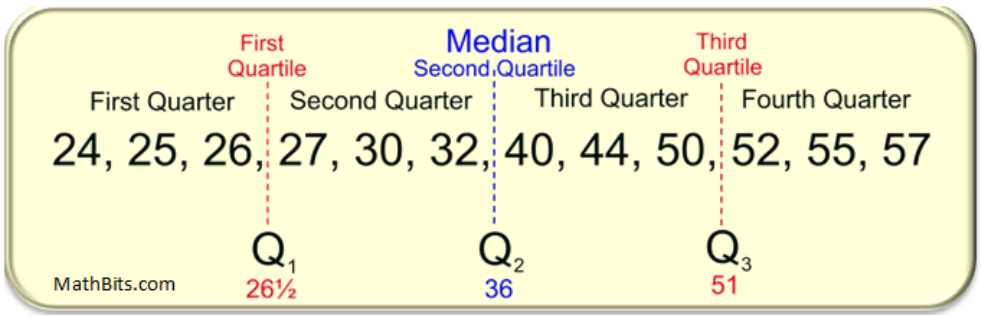
Genom att ta kvadratroten ur variansen får vi standardavvikelsen:

$$s = \sqrt{s^2} = \sqrt{36.446} = 6.037$$

Vi har räknat ut att standardavvikelsen för vikten på jordsäckarna är 6.037 kg.

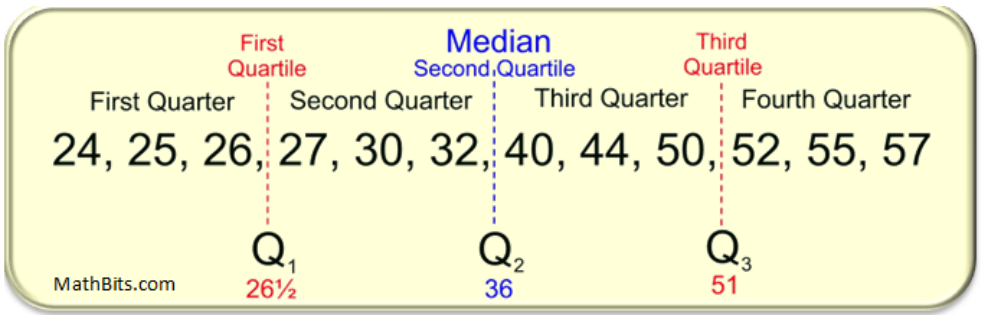
# Numeriska variabler - Kvartiler och Kvartilavstånd

- Ett annat mått på spridning är **Kvartilavstånd (interquartile range)**.
- För att förstå vad det är måste vi först förstå vad **kvartiler (quartiles)** är.
- En fördelning kan delas upp i fyra lika stora delar med hjälp av tre kvartiler, som vi kallar Q1, Q2 och Q3.
- Bilden visar värden på en variabel som sorterats i storleksordning.



# Numeriska variabler - Kvartiler och Kvartilavstånd

- **Q1:** är ett värde som är större än 25% av observationerna och mindre än de övriga 75% av observationerna.
- **Q2:** är ett värde som är större än 50% av observationerna och mindre än de övriga 50% av observationerna. Q2 är samma sak som medianen.
- **Q3:** är ett värde som är större än 75% av observationerna och mindre än de övriga 25% av observationerna.






Stockholm University

# Numeriska variabler - Kvartiler och Kvartilavstånd

Det finns ingen entydig regel för hur kvartilerna räknas ut. I De Veaux et al(2021) föreslås följande metod:

1. Sortera observationerna i storleksordning.
2. Identifiera medianen, som är samma sak som Q2.
3. Om det finns ett jämnt antal observationer så dela in dem två lika stora delar, en med lägre värden och en med högre värden. Om det är ett udda antal observationer, gör samma sak men låt observationen i mitten ingå i *båda delarna*.
4. Räkna ut medianen för observationerna med lägre värde. Detta är Q1.
5. Räkna ut medianen för observationerna med högre värde. Detta är Q3.

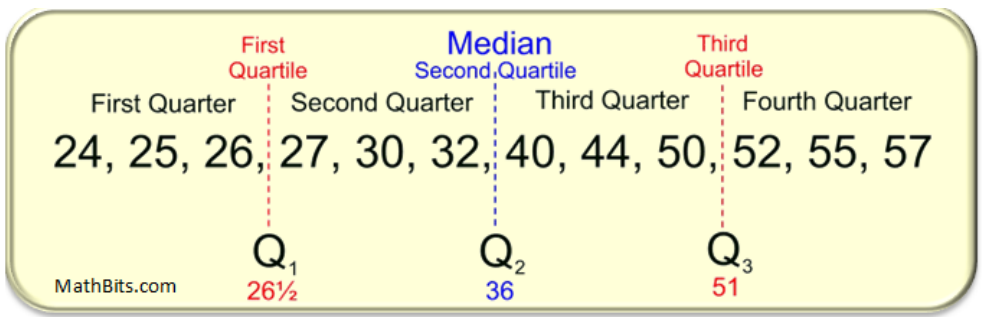
# Numeriska variabler - Kvartiler och Kvartilavstånd

Kvartilavståndet (Interquartile range) kan räkas ut som avståndet mellan Q3 och Q1.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

För fördelningen nedan kan IQR beräknas

$$\text{IQR} = \text{Q3} - \text{Q1} = 51 - 26.5 = 24.5$$



# Numeriska variabler - Kvartiler och Kvartilavstånd

- Vi kan också tala mer generellt om **percentiler**.
- Den *n:te* percentilen är ett värde som är större än *n* procent av observationerna och mindre än *100* - *n* procent av observationerna.
- **Exempel:** Den 90:e percentilen är ett värde som är större än 90 procent av observationerna och mindre än 10 procent av observationerna.
- Q1 är alltså samma sak som den 25:e percentilen, Q2 är samma sak som den 50:e percentilen och Q3 är samma sak som den 75:e percentilen.

# Numeriska variabler - Standardavvikelse eller IQR

- Om spridningen i en fördelning bäst rapporteras i form av standardavvikelse eller i form av IQR beror på syftet.
- Standardavvikelsen är bättre om det är viktigt att alla observationer beaktas.
- IQR är bättre om vi vill ha ett mått som inte påverkas av outliers.
- Standardavvikelse brukar rapporteras tillsammans med medelvärdet och IQR tillsammans med medianen.

# Funktioner i R

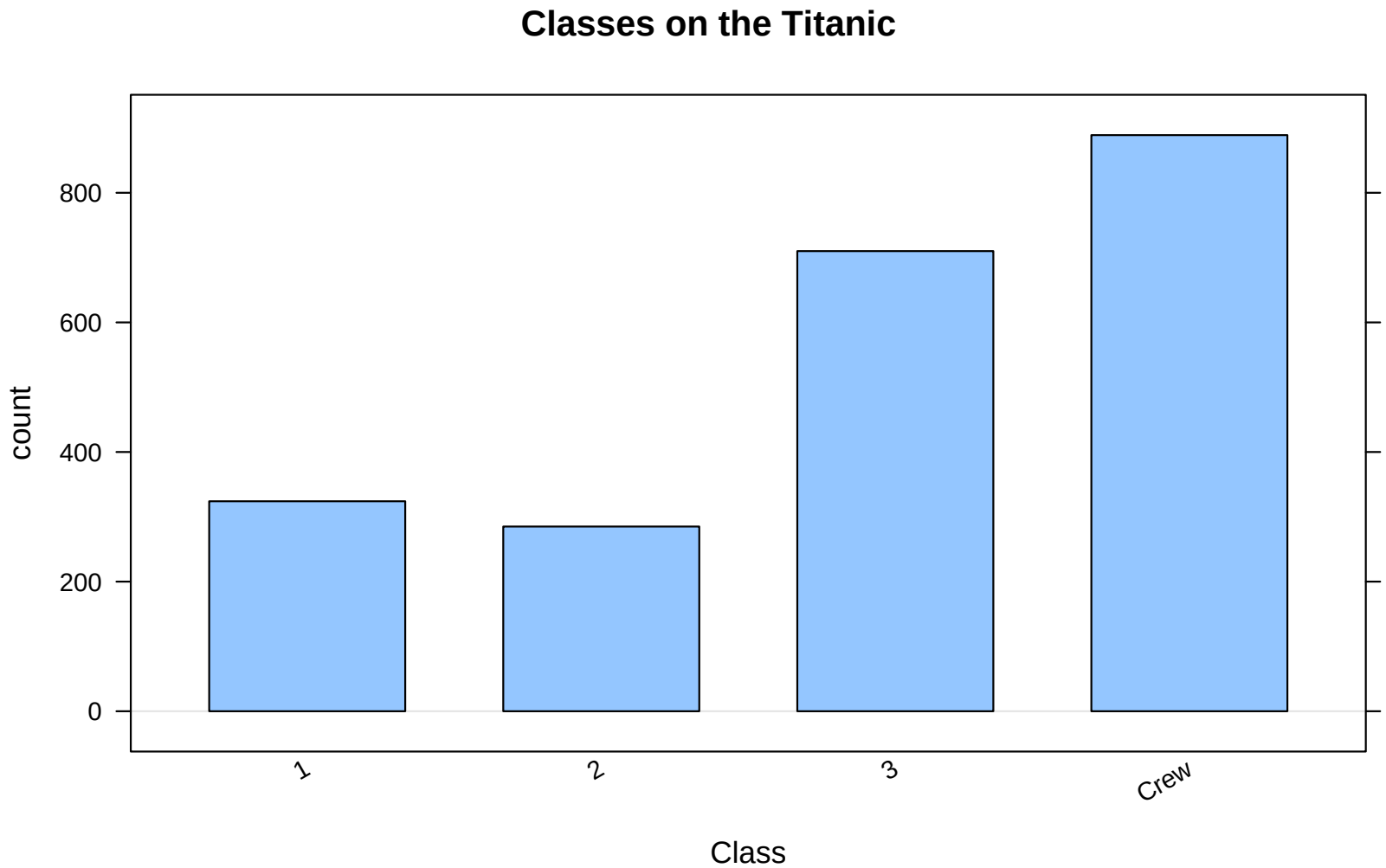
För att kunna använda den här funktionen måste du först ha läst in datasetet för Titanic. Dessutom måste du ha installerat paketet *mosaic*.

```
#Make a frequency table of variable Class
tally(~Class, data=titanic) # Requires the package mosaic
```

```
Class
   1    2    3 Crew
 324  285  710  889
```

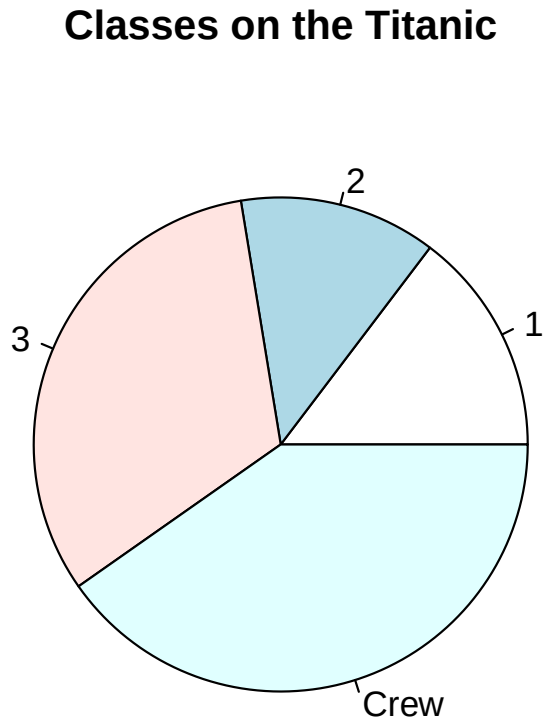
# Funktioner i R

```
#Make a barplot of variable Class
bargraph(~Class, data=titanic, main="Classes on the Titanic")
```



# Funktioner i R

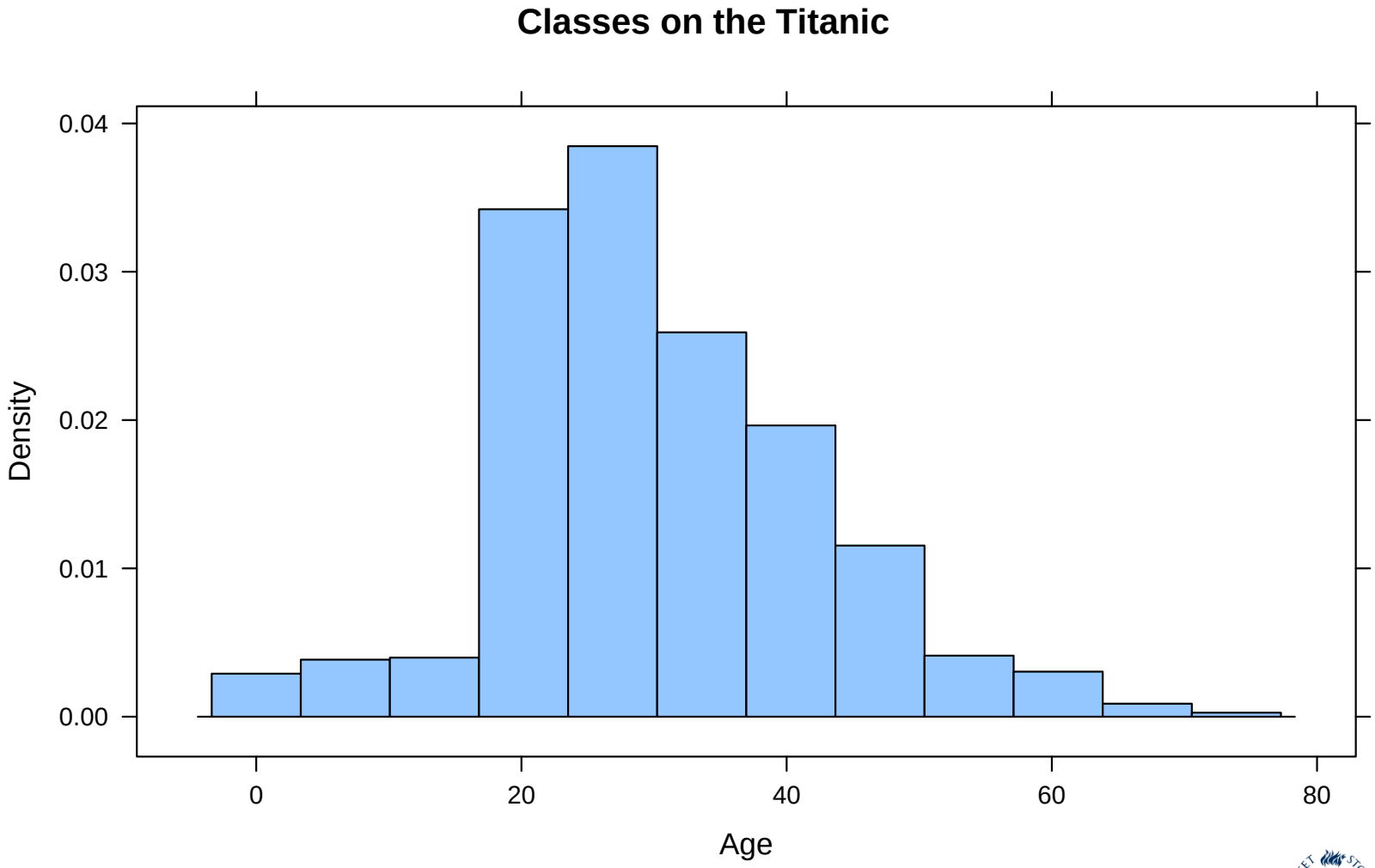
```
#Make a pie chart of the variable class
class_table <- tally(~Class, data=titanic)
pie(x=class_table, main="Classes on the Titanic")
```



# Funktioner i R

Det här kommandot ger oss ett density histogram.

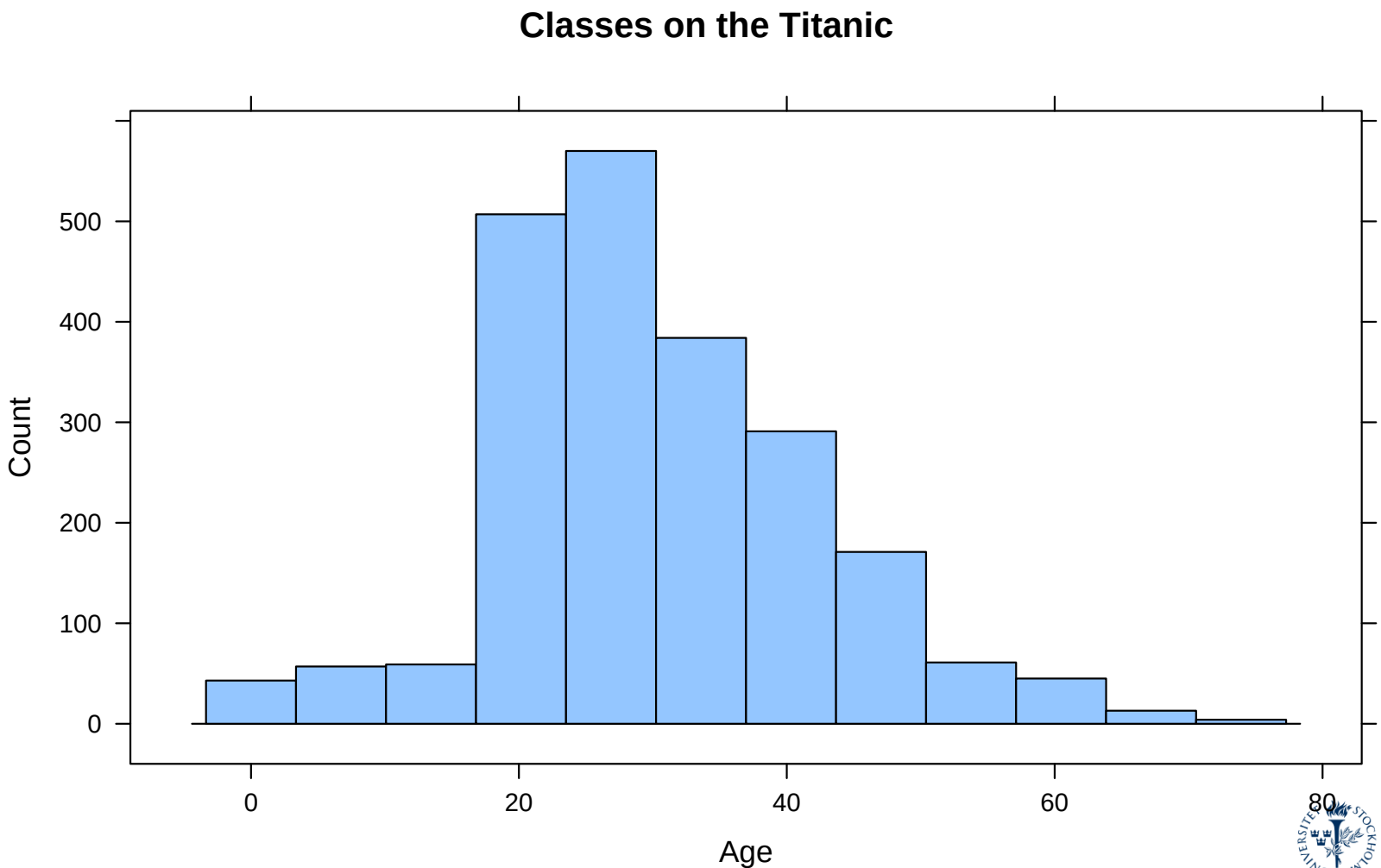
```
histogram(~Age, data=titanic, main="Classes on the Titanic")
```



# Funktioner i R

Genom att sätta type=“count” får vi ett histogram med frekvenser.

```
histogram(~Age, data=titanic, main="Classes on the Titanic",
          type="count")
```



# Funktioner i R

Funktionen favstats i mosaic ger oss flera mått som kan användas för att visa centrum och spridning i en fördelning.

Längt till höger ser vi att *missing* har värdet tre. Det betyder att tre av observationerna saknar värden för variabeln *Age*.

```
favstats(~Age, data=titanic)
```

```
  min Q1 median Q3 max     mean       sd    n missing
 0.08 22     29 37  74 30.14689 11.97386 2205       3
```