Statistik och dataanalys I

F2: Att hantera och beskriva data

Valentin Zulj

Statistiska institutionen

Stockholm
University

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
- Det du kommer lära dig här används bland annat av
  - Data scientists
  - Business analysts
  - Statistiker
  - Ekonomer
  - Forskare
  - Alla andra som behöver förstå eller förklara de insikter som ett datamaterial ger

Tips inför kursen

- Läs igenom kapitlet i boken före föreläsningen
- Om ett matematiken ser svårt ut: börja med att titta på notationen!

Exempel: För att förstå innebörden av

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

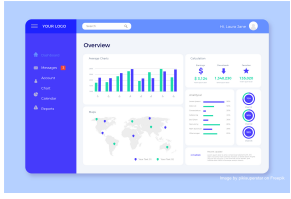
måste du först förstå vad $\bar{x}$ och n står för, och vad $\sum_{i=1}^{n} x_i$ betyder.

- Om du fastnar och inte hittar svar i boken, **fråga!**
- Skjut inte upp pluggandet, utan börja direkt!

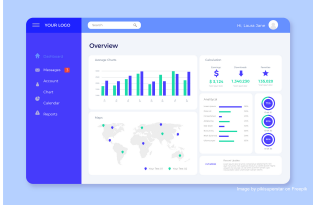
Statistik och data

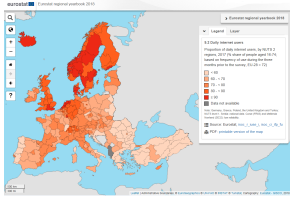
Deskriptiv statistik: Beskriv dina data på ett meningsfullt sätt

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	96.18	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	196	167
2012	Bradley Wiggins	Great Britain	32	Sky	87.58	39.83	20	3480	198	153
2013	Christopher Froome	Great Britain	28	Sky	84.55	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	3360.3	198	160
2016	Christopher Froome	Great Britain	31	Sky	89.08	39.62	21	3529	198	174
2017	Christopher Froome	Great Britain	32	Sky	86.54	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





Allting börjar med data

- Data är allt som vi kan observera och spara på ett eller annat sätt
- Data kan vara **strukturerade**

	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

- Data kan också vara **ostrukturerade**

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 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

- 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**, som är information **om** vårt datamaterial
  - *Vem* har samlat in datamaterialet
  - *Hur* är datamaterialet insamlat?
  - *När* är materialet insamlat?
  - Vad betyder variabelnamnen?
  - Hur är variablerna *kodade*?
- Metadata påverkar ofta **trovärdighet** och **användbarhet**

Olika typer av variabler

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

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

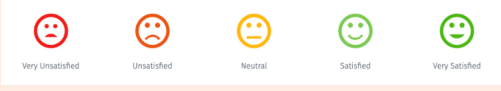
Year	Winner	Country of Origin	Age	Team	Total Time (hours)	Avg. Speed (km/h)	Stages	Total Distance Ridden (km)	Starting Riders	Finishing Riders
1903	Maurice Garin	France	32	La Française	94.55	25.7	6	2428	60	21
1904	Henri Cornet	France	20	Cycles JC	96.10	25.3	6	2428	88	23
1905	Louis Trousseller	France	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 text, men kan vara 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 är du med ditt köp, på en femgradig skala ?



- **ID-variabler** har ett unikt värde för varje observation, t.ex. personnummer i data över individer, eller årtal i data över år

Att sammanfatta kategoriska variabler

Kategoriska variabler – hur kan de beskrivas?

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

- Vi har ett klassiskt dataset om passagerare och besättning på Titanic
- Det är svårt att få en bra **överblick** med hjälp av tabellen ovan
- Vi vet att det fanns **2208** personer ombord när skeppet sjönk, men
  - Kan vi få en bra bild av antalet passagerare i varje klass?
  - Kan vi få en bild av hur överlevnad skiljer sig mellan klasser, ålder, osv?

Kategoriska variabler – hur kan de beskrivas?

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

- Vi vill sammanfatta datamaterialet, och göra det överskådligt
- Vi vill t.ex. kunna få en sammanfattning till en rapport på jobbet, eller till en uppsats/inlämning på universitetet – hur kan vi göra det?
- Tabeller och figurer!

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 (i procent) istället för antal
- Summan av andelar i den relativa frekvenstabellen ska bli 100%,

$$14.67\% + 12.91\% + 32.16\% + 40.26\% = 100\%$$

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

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

- **Notation:**
  - p_a står för andelen i procent som tillhör grupp a
  - n_a står för antalet observationer som tillhör grupp a
  - n står för det totala antalet observationer i datamaterialet

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.

Exempel: Andelen som tillhörde besättningen var

$$p_{\text{crew}} = \frac{n_{\text{crew}}}{n} \cdot 100 = \frac{889}{2208} \cdot 100 = 40.26\%,$$

där vi använt att $n = 324 + 285 + 710 + 889$

Class	Count
First	324
Second	285
Third	710
Crew	889

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

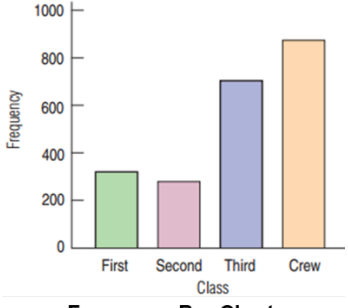
Class	Percentage (%)
First	14.67
Second	12.91
Third	32.16
Crew	40.26

Table 2.3
A relative frequency table for the same data.

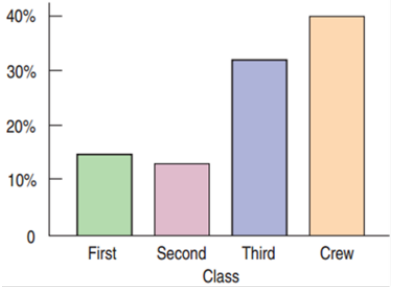
- Den här är vårt första exempel på en **fördelning** (*eng: distribution*)
- Något förenklat anger en fördelning
  - Vilka värden en variabel kan ha (*First*, *Second*, osv)
  - Hur ofta varje värde förekommer (*14.67%*, *12.91%*, osv)
- Fördelning är ett nyckelbegrepp inom statistik

- Vi kan beskriva en variabel lite 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
  - Pajdiagram (*pie chart*), till vänster
  - Stapeldiagram (*bar plot*), till höger

- Ett stapeldiagram kan vara **baserat på en frekvenstabell**, om staplarnas höjd anger **antalet** observationer som tillhör en viss grupp
- Ett stapeldiagram kan också vara **baserat på en relativ frekvenstabell**, om staplarnas höjd anger **andelen** av observationerna som tillhör en viss grupp

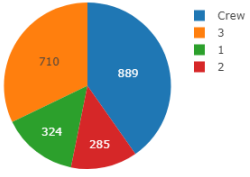


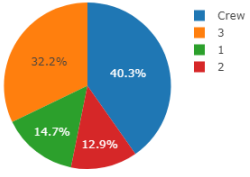
Frequency Bar Chart

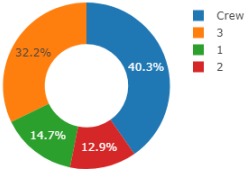


Relative Frequency Bar Chart

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

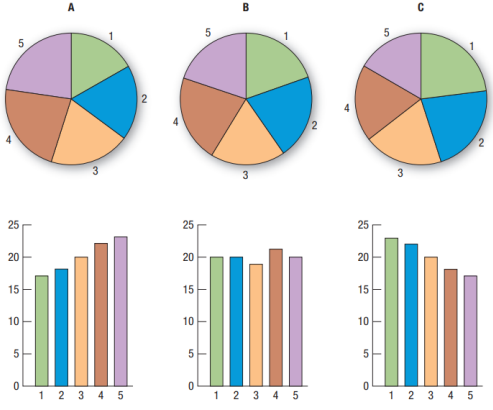






Stapeldiagram eller pajdiagram?

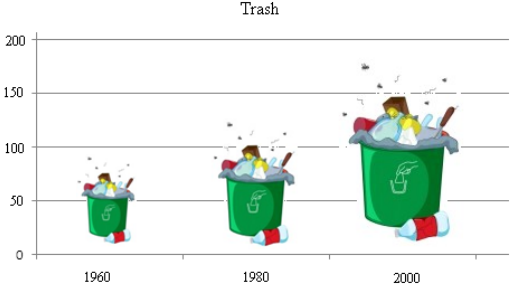
- Pajdiagram kan vara bättre om publiken har mindre erfarenhet av statistik, medan stapeldiagram brukar föredras av tekniskt kunnig publik
- I ett stapeldiagram är det lättare att se vilken grupp som är större, särskilt om staplarna står i storleksordning

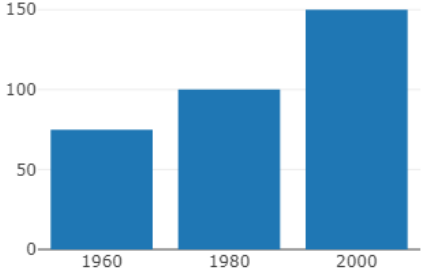


Fråga: Jämför den största soptunnan med den minsta – hur många gånger större skulle du säga att den största soptunnan är?

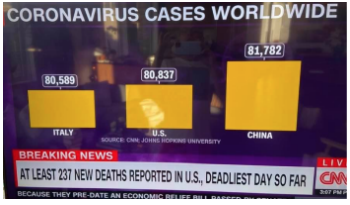
Areaprincipen

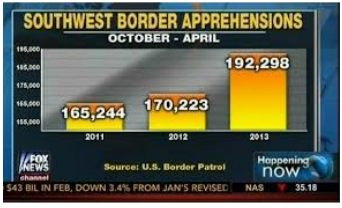
- Vi tenderar att lägga vikt vid staplarnas **area** när vi läser av ett diagram
- Arean bör så vara proprtionell mot den storlek som stapeln representerar
- Detta kallas för **areaprincipen**
- I figuren till vänster ser vi på y-axeln att den största soptupnnan är ungefär **dubbelt** så hög som den minsta, men dess area är **fyra gånger så stor**
- Stapeldiagrammet till höger ger är en mer rättvis representation



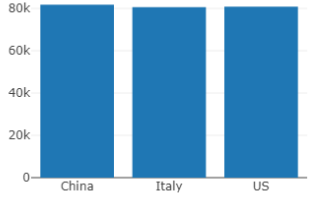


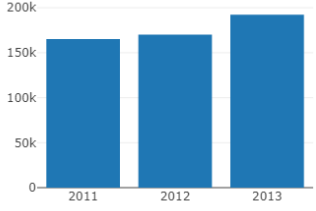
- Ibland bryter vi mot areaprincipen genom att kapa y-axeln (så att den inte börjar vid 0)





- Mer rättvisa jämförelser ges av diagrammen nedan

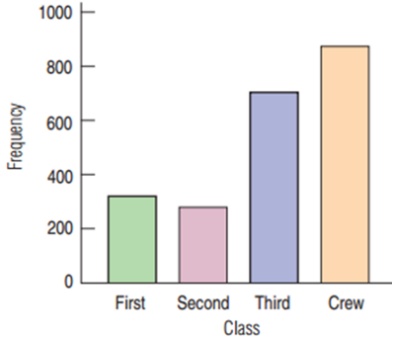


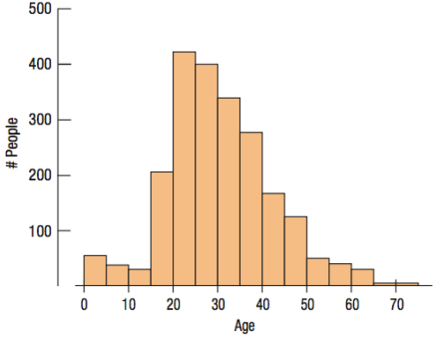


Att sammanfatta numeriska variabler

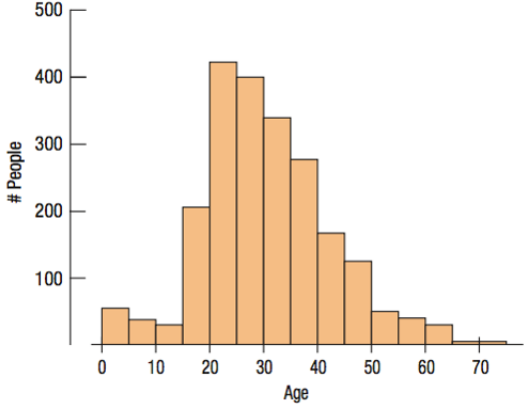
Numeriska variabler - hur kan de beskrivas?

- För numeriska variabler används vanligtvis **histogram** istället för stapeldiagram
- Histogram ser ut ungefär som stapeldiagram, men istället för kategorier representerar staplarna intervall av numeriska värden
- Till vänster ser vi ett stapeldiagram för den kategoriska variabeln *Class*
- Till höger ser vi ett histogram för den numeriska variabeln *Age*

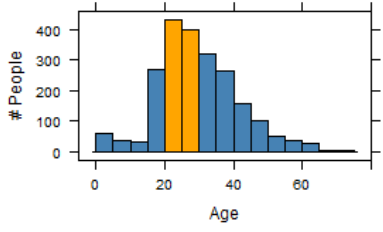


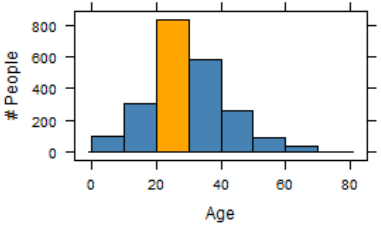


- Varje stapel representerar ett åldersintervall på 5 år
- Vi ser vi att alla passagerare på Titanic var mellan 0 och 75 år gamla
- Den fjärde stapeln från vänster visar att drygt 200 passagerare var 15-19 år



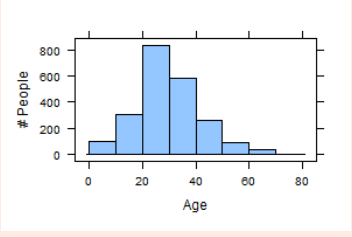
- När vi gör ett histogram väljer vi själva **bredden** på intervallen
- I histogrammet till vänster representerar de orangefärgade staplarna ungefär 400 personer vardera
- I det högra histogrammet är de vänstra intervallen ihopslagna, och den sammanslagna stapeln representerar då ungefär 800 personer

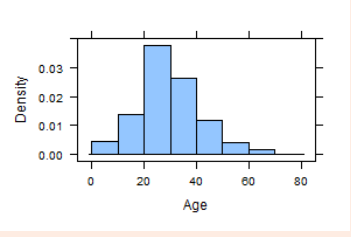




- Som alternativ finns även **täthetshistogram** (*density histogram*)
- I ett täthetshistogram motsvarar **arean** av en stapel andelen observationer som ligger inom motsvarande intervall

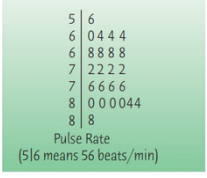
- **Exempel:**
  - Den högsta stapeln i den högra figuren har en höjd som är ≈ 0.036
  - Stapeln är 10 år bred, så arean är $0.036 \cdot 10 = 0.36$, och andelen personer i åldersintervallet 20-29 år var alltså ungefär 36%

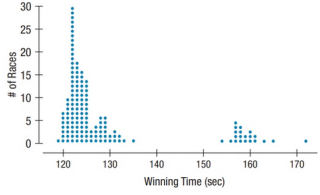




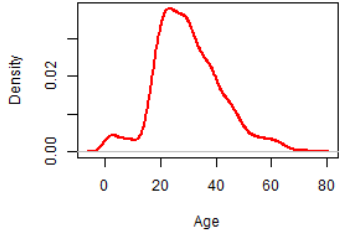
Andra typer av diagram för numeriska variabler

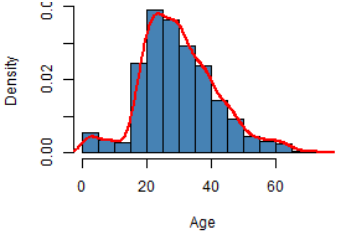
- Vi kan använda stam- och bladdiagram (vänster) och punktdiagram (höger)





- Vi kan även använda täthetsdiagram – som ett histogram fast utjämnat

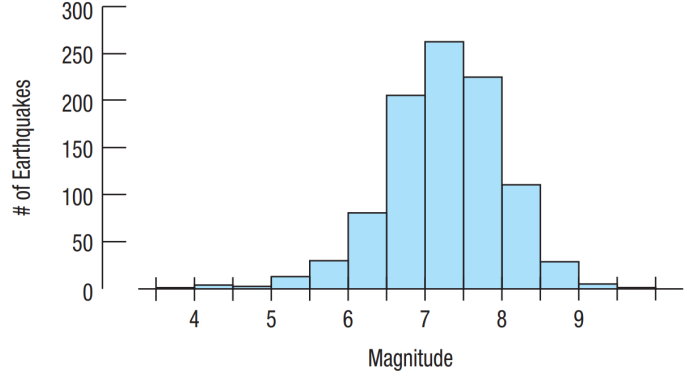




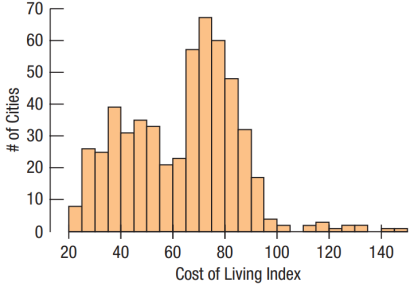
Att analysera histogram

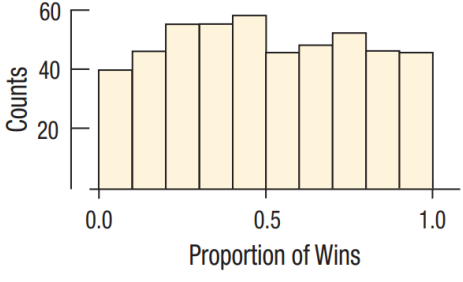
- Formen på ett histogram kan ge oss intressant information om hur värden på en variabel är fördelade
- Vi kan titta på
  - **Typvärdet** (*en: mode*) är det värde av en variabel som observerats flest gånger (det värde på x-axeln där fördelningskurvan når sin topp)
  - **Symmetrin och skevheten** (*en: symmetry/skewness*) anger om fördelningen är symmetrisk eller sned
  - **Extrema värden (outliers)** är observationer som ligger långt från övriga observationer

- Om fördelningen för en variabel har en enda topp så kallar vi fördelningen för **unimodal** (one mode) – typvärdet finns då på samma ställe som toppen
- Figuren nedan visar fördelningen magnituden på jordbävningar, och har sitt typvärde i närheten av 7

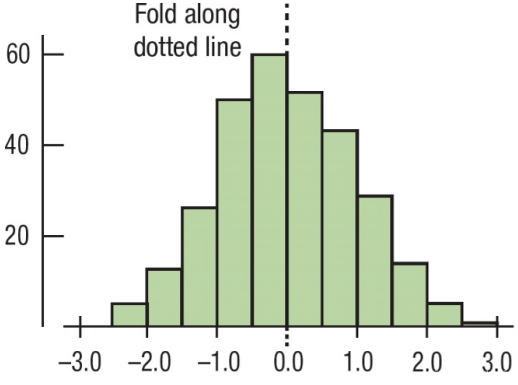


- Om en fördelning har två toppar kallas den **bimodal**, och om den har flera toppar kallas den **multimodal**
- Figuren till vänster visar ett index för levnadskostnader i olika städer, och har en topp var vid 40 och 80 (bimodal, kanske två olika grupper av städer?)
- En fördelning som är jämn utan tydliga toppar och dalar, som den till höger, kallas för en **uniform** eller **likformig fördelning**

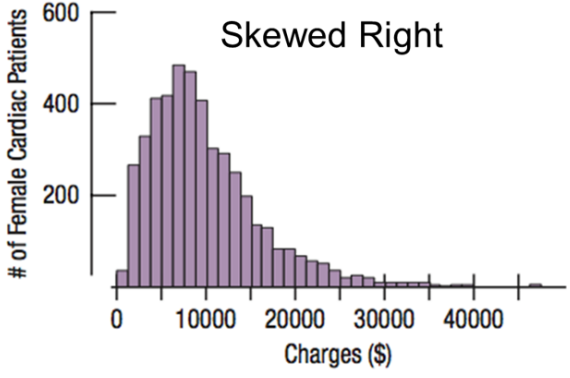




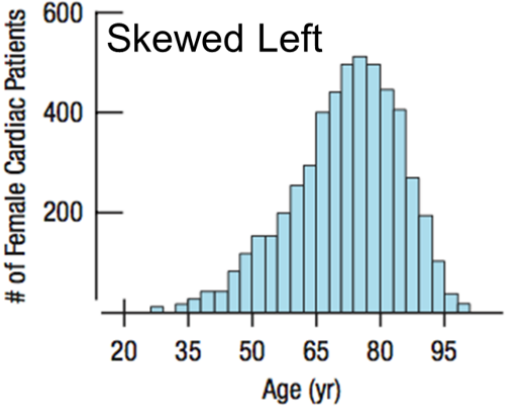
- Vi säger attt det **gröna** histogrammet är **symmetriskt**
- Den högra halvan av histogrammet är ungefär en spegelbild av den vänstra
- Denna egenskap är förvånanstvärt vanlig i naturen

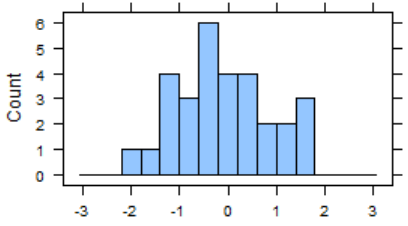


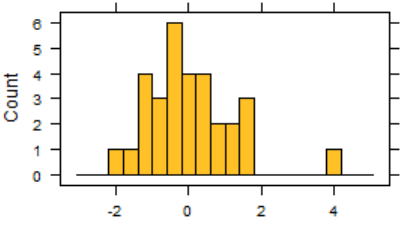
- Det **lila** histogrammet visar hur mycket kvinnliga hjärtpatienter har fakturerats för vårdbehandling
- Histogrammet är **skevt åt höger** (*right skewed*) – många patienter har betalat långt mer än typvärdet, medan få har betalat mycket mindre
- Vi säger skevt åt **höger** eftersom **högersvansen är utdragen**



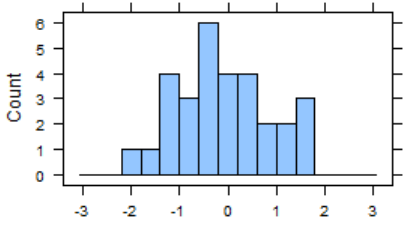
- Det **blå** histrogrammet visar åldern hos kvinnliga hjärtpatienter
- Histogrammet är **skevt åt vänster** (*left skewed*) – många patienter har en ålder långt under typvärdet, men få har en ålder långt över
- Vi säger skevt åt **vänster** eftersom **vänstersvansen är utdragen**

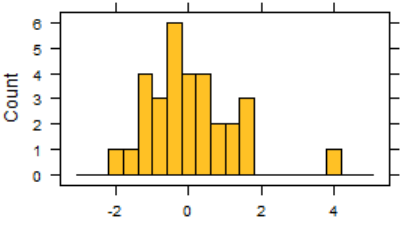






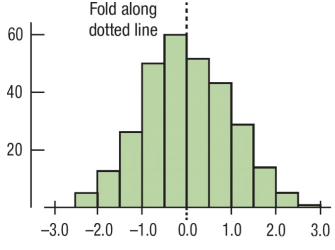
- Extrema värden som avviker från övriga observationer brukar kallas för **outliers**, även på svenska
- Det blå histogrammet har inga outliers – alla observationer är samlade nära varandra
- Det gula histogrammet har en outlier till höger om de övriga observationerna

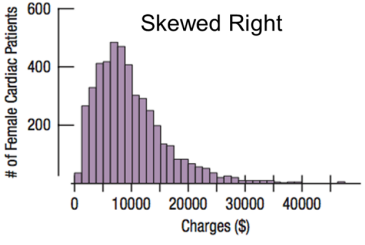




- Outliers kan få stora effekter i en statistisk analys, och behöver ofta utredas
- Outliers kan vara resultat av misstag (i mätning, inmatning, etc), men kan också vara korrekta observationer
- Om outliers tas bort ur datamaterialet måste detta dokumenteras och motiveras

- Vi vill ofta ha ett mått på det typiska värdet av en variabel, vanligtvis ett värde vid fördelningens centrum
  - I en symmetrisk fördelning finns det typiska värdet i mitten
  - I en skev fördelning är det lite svårare att ange ett rimligt typiskt värde
- Vi tar upp tre mått som alla beskriver någon sorts centrum för fördelningen: **typvärde** (*mode*), **medelvärde** (*mean*), och **median**





- Antag 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 för n observationer beräknas som

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

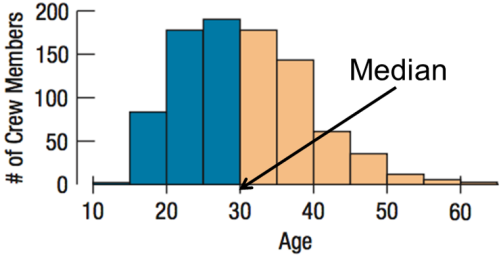
- Låt oss förklara notationen i uttrycket

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

- $\bar{x}$ (*uttalas x-streck eller x-bar*) betecknar medelvärdet för variabeln x, och motsvarande gäller för $\bar{y}$ osv
- n används som symbol för antalet observationer i våra data (i föregående exempel har vi $n = 7$)
- $\sum_{i=1}^{n} x_i$ betecknar summan av alla värden av variabeln x, dvs

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

- Medianen är ett värde som är större än ungefär hälften av observationerna och mindre än ungefär hälften av observationerna
- Vi säger *ungefär* då antalet observationer inte alltid är jämnt delbart med 2
- Figuren visar åldersfördelningen för Titanics besättning
- Vi antar att de blå staplarna motsvarar lika många personer som de i beige
- Vi har då lika många över som under 30 år, och medianåldern är ca 30



Hitta medianen

1. Sortera observationerna från lägsta till högsta värde
2. Ta fram medianen enligt nedan
   - Om antalet observationer är udda: hitta den mittersta observationen, och ange värdet på denna som median
   - Om antalet observationer är udda: identifiera de **två** observationerna som ligger i mitten, och ange deras **medelvärde** som median

Median 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

Median med udda antal observationer

- Vi har variabeln x med följande 5 värden:

x				
1.7	2.2	3.09	3.11	14.7

- Medianen är värdet i mitten av den sorterade listan

x				
1.7	2.2	3.09	3.11	14.7

- Medianen är alltså **3.09**

Median 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

Median med jämnt antal observationer

- Vi har variabeln x med följande 6 värden:

x					
1.7	2.2	3.09	3.11	14.7	16.3

- Medianen är medelvärdet av de två observationerna i mitten

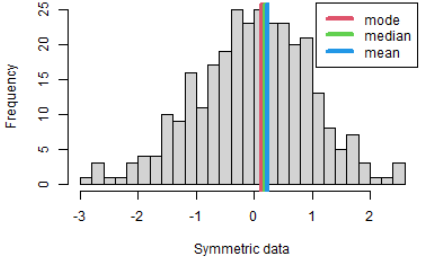
x					
1.7	2.2	3.09	3.11	14.7	16.3

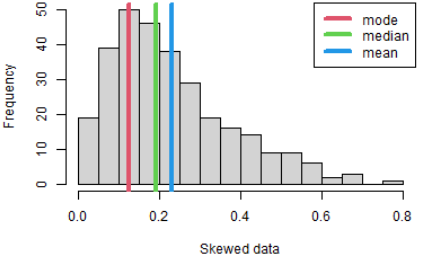
- Medianen är alltså

$$\frac{3.09 + 3.11}{2} = \mathbf{3.10}$$

Median, medelvärde eller typvärde?

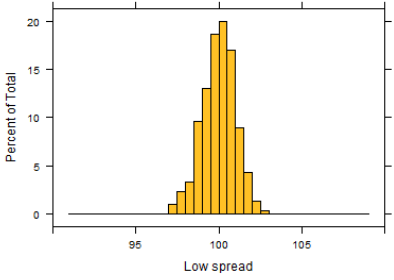
- I en symmetrisk fördelning (bild till vänster) är de tre måtten oftast snarlika
- I en skev fördelning (bild till höger) påverkas medelvärdet mer av värden ute i svansarna
- Outliers kan ha stor påverkan på medelvärdet, men inte på medianen

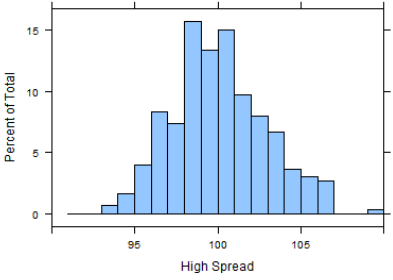




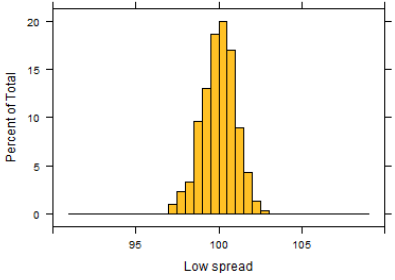
Median, medelvärde eller typvärde?

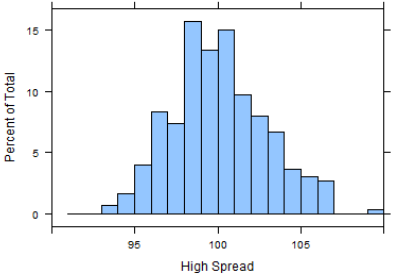
- Vilket värde som bör rapporteras beror på syftet
- Om du säljer biljetter till en båtresa och vill veta hur stora intäkterna blir är du förmodligen intresserad av medelvärdet av biljettpriset
- En köpare som undrar vad en typisk biljett kostar är kanske mer intresserad av medianpriset, som inte påverkas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





- Fördelningarna har ungefär samma medelvärde, men olika **spridning**
- Om histogrammen visar inkomstfördelningen i två länder så representerar
  - Det **gula** ett land där inkomstnivåerna är relativt lika
  - Det **blå** ett land där skillnaderna i inkomst är större





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

- **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 år
- Variationsbredden för åldersvariabeln är avståndet mellan 62 och 14, alltså

$$62 - 14 = 48$$

- **Standardavvikelsen** (betecknas s) mäter hur mycket observationerna avviker från medelvärdet
- För att hitta s är det lättast att först beräkna **variansen** (s^2),

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

- Standardavvikelsen ges sedan av kvadratroten ur variansen, alltså

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

- 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 (alltså $\bar{y}$)
- Uttrycket $\sum_{i=1}^{n}(y_i - \bar{y})^2$ är summan av dessa kvadrerade skillnader

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

- y_n är vår sista observation

- Antag att vi har mätt vikten (i kg) på nio säckar med jord

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
23	27	22	11	18	26	19	13	28

- y_1 är vikten för första säcken, y_2 är vikten för andra säcken, osv
- Som mått på spridningen av vikter vill vi räkna ut standardavvikelsen
- Vi börjar med formeln för variansen

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

- Vi behöver säckarnas medelvikt, $\bar{y}$!

- Antag att vi har mätt vikten (i kg) på nio säckar med jord

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
23	27	22	11	18	26	19	13	28

- Medelvikten beräknas som

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{23 + 27 + 22 + 11 + 18 + 26 + 19 + 13 + 28}{9} = 20.78$$

- Vi stoppar in värdet i variansformeln, och får

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - 20.78)^2}{n - 1}$$

- Antag att vi har mätt vikten (i kg) på nio säckar med jord

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
23	27	22	11	18	26	19	13	28

- Nästa steg är att beräkna de kvadrerade avvikelserna

$$
\begin{aligned}
(y_1 - 20.78)^2 &= (23 - 20.78)^2 = 4.93 \\
(y_2 - 20.78)^2 &= (27 - 20.78)^2 = 38.69 \\
&\vdots \\
(y_8 - 20.78)^2 &= (13 - 20.78)^2 = 60.53 \\
(y_9 - 20.78)^2 &= (28 - 20.78)^2 = 52.13
\end{aligned}
$$

- Antag att vi har mätt vikten (i kg) på nio säckar med jord

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
23	27	22	11	18	26	19	13	28

- Summan $\sum_{i=1}^{9}(y_i - 20.78)^2$ blir

$$4.93 + 38.69 + 1.49 + 95.65 + 7.73 + 27.25 + 3.17 + 60.53 + 52.13 = 291.57$$

- Vi stoppar in detta i formeln och får variansen

$$s^2 = \frac{\sum_{i=1}^{9}(y_i - 20.78)^2}{n-1} = \frac{291.57}{9-1} = 36.446$$

- Antag att vi har mätt vikten (i kg) på nio säckar med jord

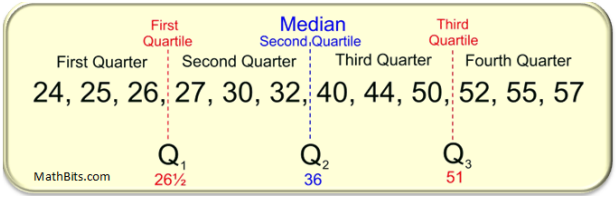
y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
23	27	22	11	18	26	19	13	28

- Till slut har vi att standardavvikelsen är

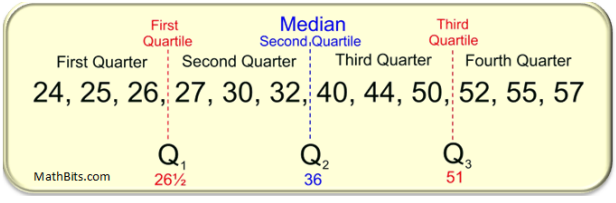
$$s = \sqrt{s^2} = \sqrt{36.446} = 6.037$$

- Vi har att **standardavvikelsen för vikten på jordsäckarna är 6.037 kg**
- **Notera:** standardavvikelsen har samma enhet (kg) som variabeln!

- 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
- För att skriva kompakt kallar vi kvartilerna för Q_1, Q_2 och Q_3
- Bilden visar värden på en variabel som sorterats i storleksordning



- Q_1: är ett värde som är större än 25% av observationerna och mindre än de övriga 75% av observationerna
- Q_2: är ett värde som är större än 50% av observationerna och mindre än de övriga 50% av observationerna (Q_2 är samma sak som medianen!)
- Q_3: är ett värde som är större än 75% av observationerna och mindre än de övriga 25% av observationerna



- Det finns ingen entydig regel för hur kvartilerna räknas ut, och i De Veaux et al (2021) föreslås följande metod:
  1. Sortera observationerna i storleksordning
  2. Identifiera medianen, som är samma sak som Q_2
     (a) Om antalet observationer är jämnt: dela in observationerna i två lika stora delar, en med mindre värden och en med större värden (än Q_2)
     (b) Om antalet observationer är udda: gör samma sak som ovan, men låt observationen i mitten ingå i *båda delarna*
  3. Räkna ut medianen för observationerna med mindre värden, detta är Q_1
  4. Räkna ut medianen för observationerna med större värden, detta är Q_3

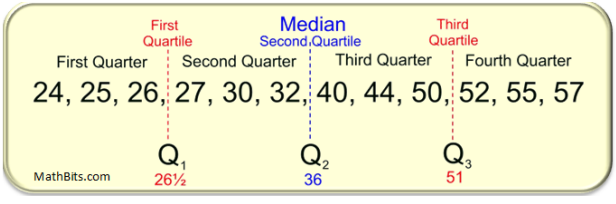
- Kvartilavståndet kan räkas ut som avståndet mellan Q_3 och Q_1

$$\text{IQR} = \text{Q}_3 - \text{Q}_1$$

- För fördelningen nedan kan IQR beräknas

$$\text{IQR} = \text{Q}_3 - \text{Q}_1 = 51 - 26.5 = 24.5$$

- Ju större kvartilavstånd, desto större spridning – observationerna är spridda över ett större intervall



- Vi kan också tala mer generellt om **percentiler**
- Den *p:te* percentilen är ett värde som är större än p procent av observationerna och mindre än $100 - p$ 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
- Q_1 är alltså samma sak som den 25:e percentilen, Q_2 är samma sak som den 50:e percentilen och Q_3 är samma sak som den 75:e percentilen

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

Statistisk programmering med hjälp av R

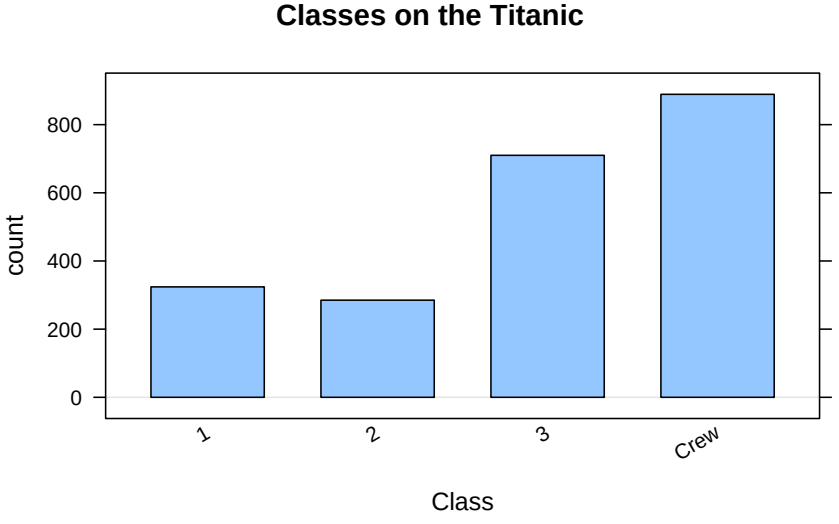
```
> library(mosaic)
> titanic <- read.csv("make_tables/Titanic.txt", sep="\t")
```

- R-koden nedan skapar en frekvenstabell som visar hur många passagerare som reste i varje klass på Titanic

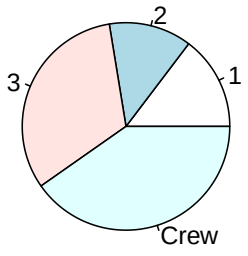
```
> #Make a frequency table of variable Class
> tally(~Class, data=titanic) # Requires the package mosaic
> Class
>    1    2    3 Crew
>  324  285  710  889
```

- För att köra koden ovan måste vi ha
  - Importerat datamaterialet `titanic` till `R`
  - Installerat paketet `mosaic`, som innehåller funktionen `tally()`

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

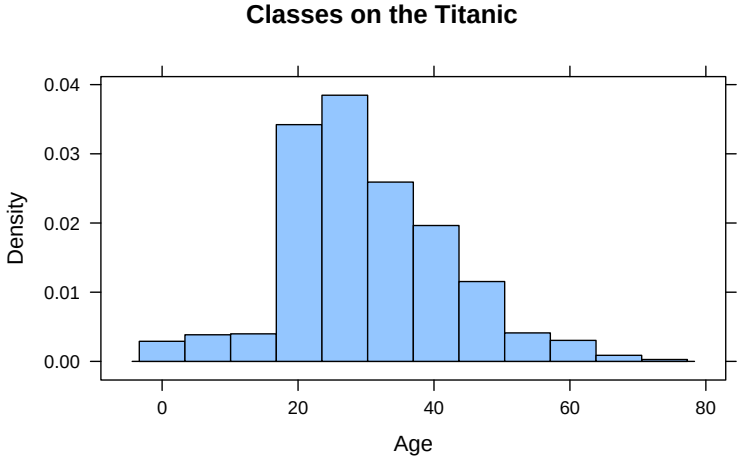


```
> #Make a pie chart of the variable class
> class_table <- tally(~Class, data=titanic) # Create freq. table
>
> # Create pie chart using freq. table
> pie(x=class_table, main="Classes on the Titanic")
```



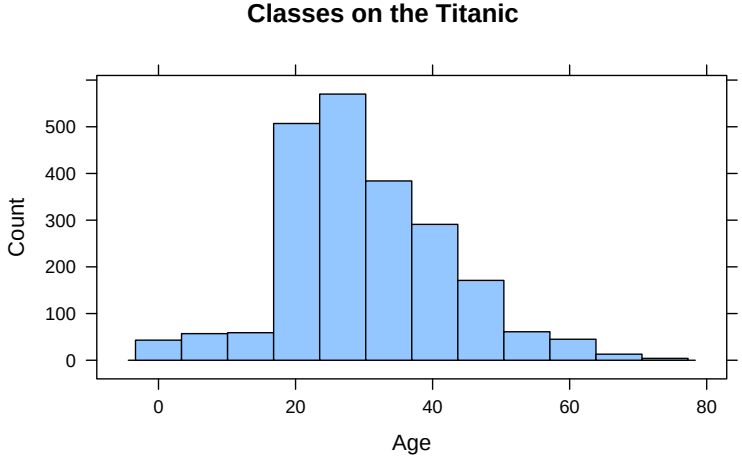
- Kommandot nedan ger oss ett **täthetshistogram**

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



- Genom att sätta `type="count"` får vi ett histogram med frekvenser

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



- 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, vilket 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
```

Tack för idag!!