

## Föreläsning 3: Hantera och beskriva data

**Matias Quiroz**<sup>1</sup>

<sup>1</sup>Statistiska institutionen, Stockholms universitet

VT 2023

- ▶ Beskriva fördelningen för kategoriska data.
- ▶ Beskriva fördelningen för numeriska data.
- ▶ Olika lägesmått och fördelningsmått.
- ▶ Olika spridningsmått.
- ▶ Deskriptiv statistik kontra inferentiell statistik.

# Data i form av tabeller

- Data kommer ofta i **tabellform**.
- Utdrag från Titanic (en färja som sjönk 14:e april 1912) datasetet:

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 Jørgensen	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

Figure 1: Tabell 2.1 i De Veaux et al. (2021).

- Datasetet innehåller en blandning av **numeriska** och **kategoriska** variabler.
- Ofta enklare att få en översikt av data genom **figurer** samt **deskriptiva sammanfattningar**.
- Figurtyp samt sammanfattningstyp avgörs av variabelns typ (numerisk eller kategorisk).

# Beskrivning av kategoriska variabler

- En **frekvenstabell** används ofta som en deskriptiv sammanfattning av kategoriska variabler. Variabeln `Class` i Titanic datasetet:

Class	Count
First	324
Second	285
Third	710
Crew	889

Figure 2: Tabell 2.2 i De Veaux et al. (2021).

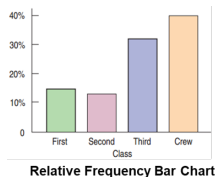
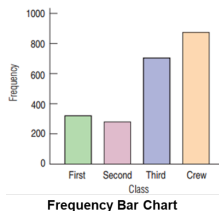
- En **relativ frekvenstabell** delar antal med totala antalet observationer. På så sätt fås **relativa frekvenser som summerar till 1**.
- Ofta anges en relativ frekvenstabell i procentform ( $\text{andel} \times 100$ ), så att cellerna i tabellen summerar till 100%.

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

Figure 3: Tabell 2.2 i De Veaux et al. (2021).

# Beskrivning av kategoriska variabler, forts.

- ▶ För att ytterligare överskådliggöra data kan man göra **grafiska beskrivningar** av en kategorisk variabel:
  - ▶ Stapeldiagram (**bar plot/bar chart** på engelska).
  - ▶ Pajdiagram (**pie chart** på engelska).
  - ▶ Munkdiagram (**donut chart** på engelska).
- ▶ *“En bild säger mer än tusen ord”*. Gör alltid en grafisk beskrivning av variablerna du vill analysera (när så möjligt).
- ▶ **Stapeldiagram** för variabeln `Class` i Titanic datasetet:



**Figure 4:** Figur 2.2 i De Veaux et al. (2021). Baserat på frekvenstabellen (vänster) och relativa frekvenstabellen (höger).

# Beskrivning av kategoriska variabler, forts.

- ▶ Enkelt att hitta data över aktuella händelser på webben.
- ▶ Exempelvis ger en Google sökning på “Antal döda covid Sverige” en länk till Socialstyrelsens hemsida.
- ▶ Antal döda uppdelade på ålderskategori samt kön (11 januari 2023):

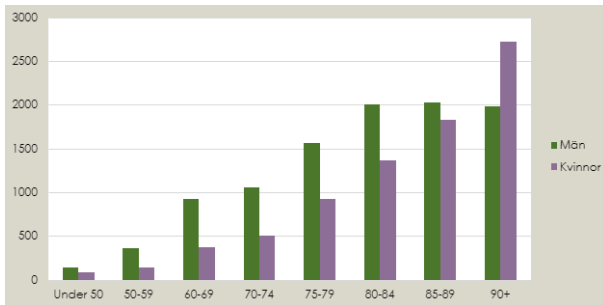


Figure 5: Källa: [www.socialstyrelsen.se](http://www.socialstyrelsen.se)

## Beskrivning av kategoriska variabler, forts.

- ▶ Ett stapeldiagram visar **variabelns fördelning** över dess olika möjliga utfall.
- ▶ Statistik handlar om att modellera data (beskriva data) med hjälp av **teoretiska fördelningar**. Teoretiska fördelningar kommer i andra delen av kursen (behövs lite sannolikhetslära först).
- ▶ Ett stapeldiagram kan ses som en **empirisk version** (räknad på observerade data) av den teoretiska fördelningen för en kategorisk variabel.
- ▶ Ett **pajdiagram** är ett annat sätt att illustrera samma information som i ett stapeldiagram (notera annorlunda färger än tidigare).

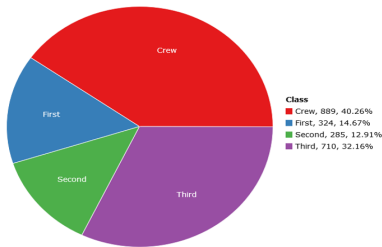


Figure 6: Från lärmaterialet skapat av utgivaren av De Veaux et al. (2021).

# Beskrivning av kategoriska variabler, forts.

- ▶ Ett **munkdiagram** är ett pajdiagram med ett hål i mitten, dvs innehåller endast pajskalet.

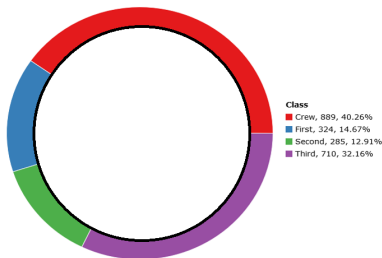


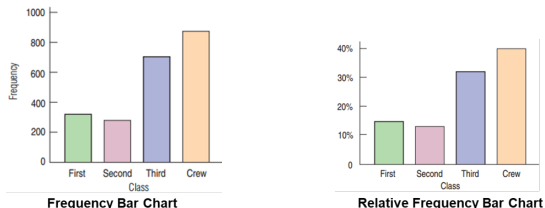
Figure 7: Återskapad från Figur 5 i föregående slide.

- ▶ Varför munkdiagram? Anses vara visuellt enklare att tyda av vissa.
- ▶ Paj eller munkdiagram kan vara fördelaktiga när man presenterar för icke-statistiker (t.ex för en styrelse eller för ledningen i ett företag).
- ▶ Stapeldiagram är att föredra när man presenterar för en teknisk kunnig publik.



# Beskrivning av kategoriska variabler, forts.

- ▶ Stapeldiagram kräver lite mer design, då det är viktigt att **areaprincipen** bevaras.
- ▶ Areaprincipen: Stapelarean ska motsvara det faktiska numeriska värdet. Exempel: Om utfall A förekommer tre gånger oftare än utfall B, så ska dess stapelarea motsvara 3 gånger stapelarean för utfall B.
- ▶ Areaprincipen var uppfylld i de tidigare stapeldiagrammen vi såg:



**Figure 8:** Figur 2.2 i De Veaux et al. (2021). Baserat på frekvenstabellen (vänster) och relativa frekvenstabellen (höger).

# Beskrivning av kategoriska variabler, forts.

- Att inte respektera areaprincipen ger visuellt misledande resultat:



Figure 9: Från @ndrew\_lawrences Twitter profil.

- Pajdiagram har fördelen att per definition respektera areaprincipen...
- ... eller?

# Beskrivning av numeriska variabler

- ▶ För **numeriska variabler**, också kallade **kvantitativa variabler**, kan vi inte göra stapeldiagram eftersom de är inte har kategorier.
- ▶ Motsvarigheten till stapeldiagram för numeriska variabler kallas för **histogram**.
- ▶ I ett stapeldiagram kan utfallen ordnas godtyckligt, eftersom **kategorier inte har en naturlig ordning**<sup>1</sup>.
- ▶ **Kvantitativa variabler har en naturlig ordning**, eftersom de innehåller numeriska värden.
- ▶ De numeriska värdena **sorteras i stigande ordning**.
- ▶ Låt oss göra ett histogram på variabeln ålder (Age) i Titanic datasetet.

---

<sup>1</sup>Sant för kategoriska variabler på **nominal skala**. Kategoriska variabler på **ordinal skala** har en naturlig ordning. Exempel: Hur ofta tränar du (sällan eller aldrig/ibland/ofta/väldigt ofta)?

# Beskrivning av numeriska variabler, forts.

- Ett histogram baserat på 15 **grupper** med **intervalllängd** 5 år:

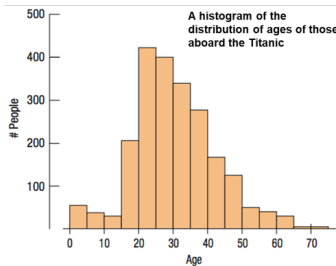


Figure 10: Från lärmaterialet skapat av utgivaren av De Veaux et al. (2021).

- Höjden av en stapel avgörs av hur många observationer som fanns i motsvarande intervall.
- Histogrammet visar att flest passagerare fanns i åldersintervallet 20-24.
- Histogrammet visar att åldersfördelningen avtar efter 25 år, och att det fanns fler spädbarn och barn än barn i de yngre tonåren.

# Beskrivning av numeriska variabler, forts.

- ▶ Histogrammets utseende beror på antal grupper (och därmed intervallängden).
- ▶ Det finns tumregler för att välja antal grupper (bins på engelska).
- ▶ En sådan tumregel är **Sturges regel**, som lyder

$$\#bins = 1 + \text{ceil}(\log_2(n)),$$

där  $n$  är antalet observationer,  $\log_2$  är logaritmfunktionen med basen 2, och  $\text{ceil}$  är en funktion som avrundar uppåt till närmaste heltal.

- ▶ Det finns andra tumregler som också tar **hänsyn till spridningen av data**, och inte bara antalet observationer.

# Beskrivning av numeriska variabler, forts.

- ▶ Vi har konstaterat att histogrammets utseende beror på antal grupper (och därmed intervalllängden).
- ▶ Viktigt att ibland testa olika antal grupper, för att se om extra information kan erhållas.
- ▶ 1116 jordbävningsmagnituder från 426 f.Kr.–2018:

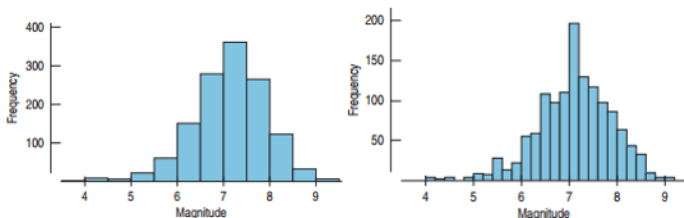


Figure 11: Från lärmaterialet skapat av utgivaren av De Veaux et al. (2021).

- ▶ Fler grupper visar en mycket tydligare “topp” på magnituder strax över 7.

# Beskrivning av numeriska variabler, forts.

- En täthetsplot (**density plot** på engelska) är ett alternativ till histogrammet vars syfte är att **reducera effekten av antalet grupper**.
- Genom att utjämna (**smooth** på engelska) staplarna reduceras effekten av antalet grupper.

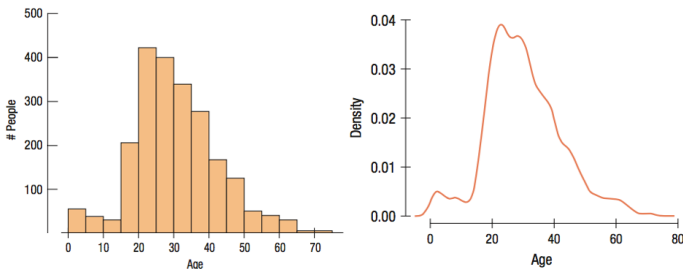
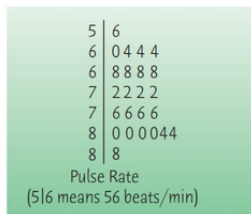
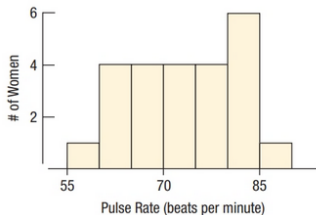


Figure 12: Figur 2.7 och 2.10 i De Veaux et al. (2021). Histogram (vänster) och täthetsplot (höger).

# Beskrivning av numeriska variabler, forts.

- ▶ Histogrammet döljer viss information: Vi vet inte hur värdena är fördelade inom varje grupp.
- ▶ När man inte har mycket data finns det andra figurer som ger ytterligare information.
- ▶ Stam och blad diagram (**stem-and-leaf display** på engelska) till höger:



**Figure 13:** Figur 2.8 i De Veaux et al. (2021). Histogram (vänster) samt stam och blad diagram (höger).

- ▶ Stam och blad diagrammet visar hur de enskilda värdena är fördelade inom varje grupp.



# Beskrivning av numeriska variabler, forts.

- ▶ En punktplot (**dotplot** på engelska) är ett modernare alternativ till stam och blad diagrammet.
- ▶ Segrartider (sekunder) i Kentucky Derby (hästkapplöpning i USA) under 1875–2018:

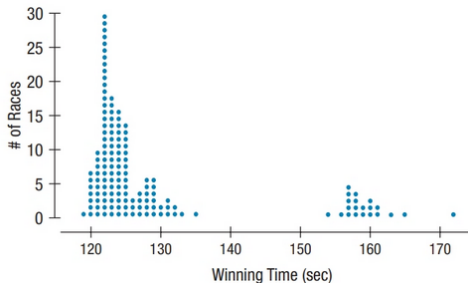


Figure 14: Figur 2.9 i De Veaux et al. (2021).

- ▶ Likt stam och blad diagrammet, så **visar punktplotten en mer detaljerad fördelning än histogrammet.**

# Beskriva fördelningens form för numeriska variabler

- ▶ Kom ihåg: Statistik handlar om att modellera data (beskriva data) med hjälp av **teoretiska fördelningar**.
- ▶ I fallet med kategoriska data så kunde ett stapeldiagram ses som en **empirisk version** (räknad på observerade data) av den teoretiska fördelningen för en kategorisk variabel.
- ▶ På samma sätt, så kan ett histogram eller täthetsplot ses som en **empirisk version** (räknad med hjälp av observerade data) av den teoretiska fördelningen för en **numerisk variabel**.
- ▶ Det är av yttersta vikt att beskriva formen (**shape** på engelska) som datafördelningen visar.
- ▶ De följande tre attribut används för att beskriva fördelningens form:
  1. Typvärde/Typvärdena (**mode/modes** på engelska) av fördelningen.
  2. Fördelningens symmetri eller skevhet (**symmetry** eller **skewness** på engelska).
  3. Fördelningens extrema värden (**outliers** på engelska).

# Beskriva fördelningens form för numeriska variabler, forts.

- ▶ Ett exempel på en variabel med **unimodal** fördelning är magnitudvariabeln i jordbävningsexemplet vi såg tidigare.
- ▶ Ett exempel på en variabel med **bimodal** fördelning är Cost of Living Index.

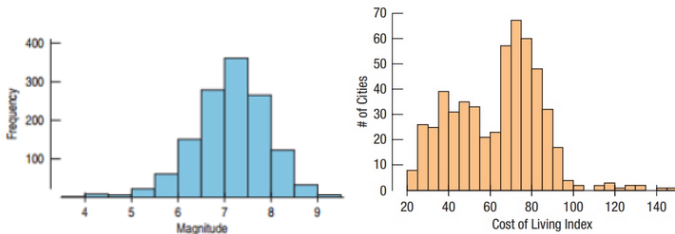


Figure 15: Från lärmaterialet skapat av utgivaren av De Veaux et al. (2021) (vänster) och på s.56 i De Veaux et al. (2021) (höger). Magnitud (vänster) och Cost of Living (höger).

- ▶ Fler än två typvärden kan finnas och sådana fördelningar kallas för **multimodala**.

# Beskriva fördelningens form för numeriska variabler, forts.

- ▶ En fördelning där staplarna är lika höga kallas en likformig (**uniform** på engelska).
- ▶ Ett exempel på en variabel med (approximativ) likformig fördelning:

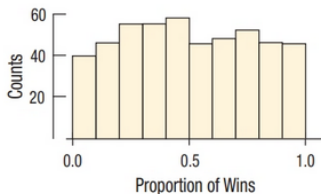


Figure 16: Figur 2.11 i De Veaux et al. (2021).

- ▶ Notera att även om det verkar finnas ett typvärde strax under 0.5, så är den inte tydlig som i de andra figurerna.
- ▶ Hur hade histogrammet sett ut **om vi hade samlat in nya data?**
- ▶ Den underliggande teoretiska fördelningen är uniform, men på grund av variation när vi tar ett stickprov ser vi små avvikelser.

# Beskriva fördelningens form för numeriska variabler, forts.

- Fördelningens (approximativa) symmetri kan avgöras genom att **spegla fördelningen runt mitten**.
- Ett exempel på en symmetrisk variabel:

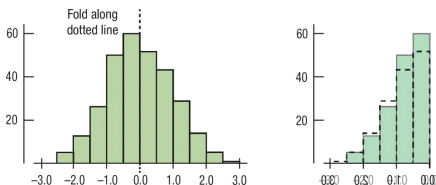


Figure 17: Från lärmaterialet skapat av utgivaren av De Veaux et al. (2021).

- Fördelningens form ovan påminner om en kyrkklocka (**bell-shaped distribution** på engelska).
- **Normalfördelningen** kommer användas senare i kursen för att beskriva sådana symmetriska fördelningar.

# Beskriva fördelningens form för numeriska variabler, forts.

- En variabel kan ha symmetrisk fördelning även om den inte är formad som en kyrkklocka.

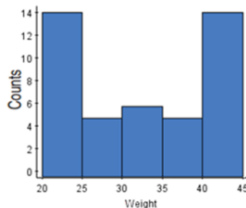


Figure 18: Från lärmaterialet skapat av utgivaren av De Veaux et al. (2021).

- Skeva fördelningar (**skewed distributions** på engelska) är exempel på fördelningar som inte är symmetriska.

# Beskriva fördelningens form för numeriska variabler, forts.

- ▶ En fördelning är negativt skev/skev åt vänster (**skewed to the left** på engelska) om fördelningen har en längre svans (**tail** på engelska) åt vänster.
- ▶ En fördelning är positivt skev/skev åt höger (**skewed to the right** på engelska) om fördelningen har en längre svans åt höger.
- ▶ Exempel på skeva fördelningar:

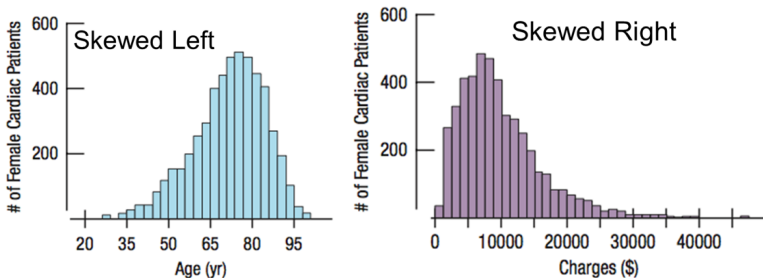


Figure 19: Från lärmaterialet skapat av utgivaren av De Veaux et al. (2021).

# Beskriva fördelningens form för numeriska variabler, forts.

- ▶ Det sista attributet för att beskriva fördelningens form är **extrema värden**.
- ▶ Problemet med extrema värden är att de kan ha en stor inverkan på våra statistiska resultat.
- ▶ Extrema värden kan förekomma **på grund av misstag**. De kan då tas bort med gott samvete.
- ▶ Extrema värden kan också förekomma **på grund av naturlig variation**.
- ▶ Exempel på extremvärden upkomna ur naturlig variation:
  - ▶ Isaac Newtons IQ (uppskattad till 190).
  - ▶ Elon Musks nettoförmögenhet (uppskattad till 203 miljarder USD 2022, efter köpet av Twitter).
- ▶ Viktigt att rapportera extremvärden som förekommer på grund av naturlig variation.



# Fördelningens centrum för numeriska variabler, forts.

- ▶ Vi har beskrivit fördelningens form genom följande tre attribut:
  1. Typvärde/typvärden (**mode/modes** på engelska) av fördelningen.
  2. Fördelningens symmetri eller skevhet (**symmetry** eller **skewness** på engelska).
  3. Fördelningens extrema värden (**outliers** på engelska).
- ▶ En annan viktig aspekt av en fördelning är dess centrala värde (**typical value** på engelska).
- ▶ Intuitivt så kommer ett centralt värde att finnas någonstans runt fördelningens centrum (**center** på engelska).
- ▶ Vi kan tänka oss olika mått som beskriver ett centralt värde:
  1. Typvärdet.
  2. Median (**median** på engelska).
  3. Medelvärde (**mean** på engelska).
- ▶ Dessa är lägesmått då de beskriver vart fördelningens centrum är lokaliserad.

# Fördelningens centrum för numeriska variabler, forts.

- ▶ Låt  $y_1, y_2, \dots, y_n$  beteckna ett stickprov (**sample** på engelska) av storlek  $n$ .
- ▶ Ett stickprov innehåller  $n$  observerade värden från en variabel ( $y$ )<sup>2</sup>.
- ▶ Medianen **delar stickprovet i två lika stora delar**: Hälften av observationerna ligger till vänster om medianen, och den andra halvan till höger om medianen.
- ▶ För att beräkna medianen måste vi först ordna stickprovet (från minsta värdet till största värdet) och sedan dela det på mitten.
- ▶ Exempel: Antag att stickprovet är 14.7, 2.2, 1.7, 3.09, 3.11, med  $n = 5$ .
  - ▶ Det ordnade stickprovet är 1.7, 2.2, **3.09**, 3.11, 14.7.
  - ▶ Värdet 3.09 delar stickprovet i två lika stora delar (2 observationer är  $< 3.09$  och 2 observationer är  $> 3.09$ ) och är således medianen.
- ▶ Vad händer om vi istället har  $n = 6$  observationer, där den sjätte observationen är 16.3?

---

<sup>2</sup>Ett stickprov kan också innehålla observerade värden från flera variabler.

# Fördelningens centrum för numeriska variabler, forts.

- ▶ Om vi har ett jämnt antal värden, t.ex  $n = 6$ , så finns det inte en enskild observation bland stickprovet som delar stickprovet i två lika stora delar.
- ▶ Då  $n$  är ett jämnt tal definieras medianen istället som medelvärdet av de mittersta två observationer. I vårt exempel  $(3.09 + 3.11)/2 = 3.10$ .
- ▶ Medianen för åldersvariabeln i Titanic datasetet är 30:

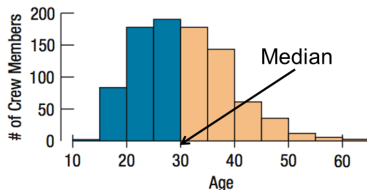


Figure 20: Från lärmaterialet skapat av utgivaren av De Veaux et al. (2021).

- ▶ Om fördelningen är (perfekt) symmetrisk så kommer medianen att återfinnas exakt i mitten.

# Fördelningens centrum för numeriska variabler, forts.

- ▶ Ett annat mått på fördelningens centrum är det **aritmetiska medelvärdet**, ofta enbart kallat medelvärdet.
- ▶ Medelvärdet  $\bar{y}$  defineras som

$$\bar{y} = \frac{\text{Totalen}}{n} = \frac{\sum y}{n}.$$

- ▶ Den grekiska symbolen  $\sum$  står för **summera**.
- ▶ Kursboken använder notationen  $\sum y$  och det är underförstått att man summerar över alla  $y$  i stickprovet.
- ▶ Andra böcker använder istället

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n},$$

för att tydligare illustrera att man tar summan av stickprovet  $y_1, y_2, \dots, y_n$ .

- ▶ Vi följer bokens konvention så långt som möjligt, men det är **viktigt att förstå båda notationerna**.

# Fördelningens centrum för numeriska variabler, forts.

- ▶ Medianen angav det värde som delar fördelningen på mitten, dvs **fördelningens mittpunkt**.
- ▶ Medianen påverkas inte av vad som händer ute i fördelningens svansar. **Medianen påverkas således inte av extremvärden**.
- ▶ Medelvärde anger istället fördelningens tyngdpunkt (**balancing point** på engelska).

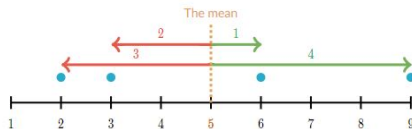


Figure 21: Källa: [khanacademy.org](https://www.khanacademy.org).

- ▶ Notera att summan av avstånden från varje punkt till medelvärdet på vardera sida är densamma (5 i det här fallet).
- ▶ Påverkas medelvärdet av extrema observationer?

# Fördelningens centrum för numeriska variabler, forts.

- ▶ Ja, tyngdpunkten kommer att flyttas. Till skillnad från medianen, **påverkas medelvärdet av alla observationer** på grund av  $\sum y$ .
- ▶ Exempel: Antag att stickprovet är 14.7, 2.2, 1.7, 3.09, 3.11, med  $n = 5$ . Då är

$$\bar{y} = \frac{\sum y}{n} = \frac{14.7 + 2.2 + 1.7 + 3.09 + 3.11}{5} = 4.96.$$

- ▶ Medelvärdet av åldersvariabeln för personalen i Titanic datasetet:

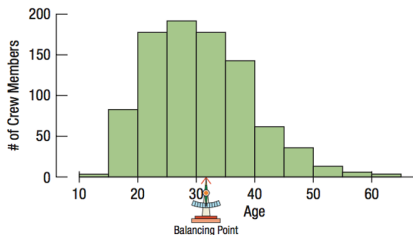


Figure 22: Figur 2.16 i De Veaux et al. (2021).

- ▶ Ofta bra att använda både medelvärdet och medianen. I del två av kursen kommer vi se att medelvärdet har **enklare statistiska egenskaper**.

# Fördelningens spridning för numeriska variabler

- ▶ Snabb repetition på fördelningens egenskaper vi har gått igenom.
- ▶ Vi har beskrivit fördelningens form genom följande tre attribut:
  1. Typvärde/Typvärden (**mode/modes** på engelska) av fördelningen.
  2. Fördelningens **symmetri** eller skevhet (**skewness** på engelska).
  3. Fördelningens extrema värden (**outliers** på engelska).
- ▶ Vi har beskrivit fördelningens centrum genom följande två lägesmått.
  1. Median (**median** på engelska).
  2. Medelvärde (**mean** på engelska).
- ▶ En sista mycket viktig aspekt av en fördelning är dess spridning/variation (**spread/variation** på engelska).
- ▶ Vi kan tänka oss olika mått som beskriver spridningen:
  1. Variationsbredd (**range** på engelska).
  2. Kvartilavstånd (**interquartile range** på engelska).
  3. Standardavvikelse (**standard deviation** på engelska).

# Fördelningens spridning för numeriska variabler, forts.

- ▶ Mått som beskriver spridningen kallas för spridningsmått.
- ▶ Fördelningens variationsbredd definieras som skillnaden mellan det största värdet och det minsta värdet i stickprovet.
- ▶ Exempel: Bland Titanics personal var den äldsta 62 och den yngsta 15, således är variationsbredden  $62 - 14 = 48$  år.
- ▶ Variationsbredden är inte så informativ angående spridningen eftersom den **ignorerar hur fördelningen ser ut mellan minsta och största värdet**.
- ▶ Dessutom **påverkas variationsbredden enormt av extremvärden**.
- ▶ För att definiera kvartilavstånd måste vi först definera kvartiler (**quartiles** på engelska).



# Fördelningens spridning för numeriska variabler, forts.

- En fördelning delas upp i fyra (quarto betyder fjärdedel på latin) lika stora bitar med hjälp av tre kvartiler.
  1. **Q1**: den observationen som har 25% av observationerna till vänster om sig (och 75% till höger om sig). Kallas även 25% percentilen (**percentile** på engelska).
  2. **Q2**: den observationen som har 50% av observationerna till vänster om sig (och 50% till höger om sig). Kallas även 50% percentilen.
  3. **Q3**: den observationen som har 75% av observationerna till vänster om sig (och 25% till höger om sig). Kallas även 75% percentilen.

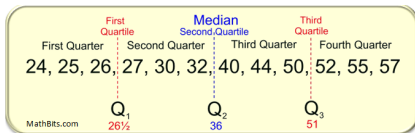


Figure 23: Källa: Mathbits.com.

- Vi kallar kvartiler för fördelningsmått.
- Känner vi igen Q2 (50% percentilen)?

# Fördelningens spridning för numeriska variabler, forts.

- ▶ Q2 är medianen! Delar upp fördelningen i två lika stora delar.
- ▶ Kvartilavståndet, förkortat **IQR** (interquartile range) definieras som avståndet mellan övre kvartilen (Q3) och undre kvartilen (Q1)

$$IQR = Q3 - Q1.$$

- ▶ Kvartilavståndet för åldersvariabeln för personalen i Titanic datasetet:

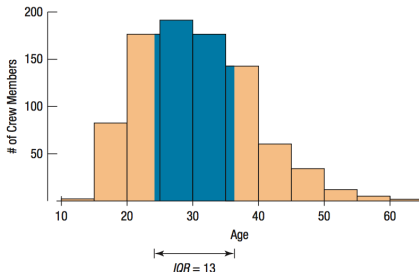


Figure 24: Figur 2.18 i De Veaux et al. (2021).

## Fördelningens spridning för numeriska variabler, forts.

- ▶ IQR påverkas inte av extremvärden.
- ▶ Som spridningsmått ignorerar IQR mycket av data.
- ▶ Dessutom tar den inte hänsyn till spridningen inom varje kvartil.
- ▶ Ett bättre spridningsmått som tar hänsyn till all data är standardavvikelsen (**standard deviation** på engelska).
- ▶ För att definiera standardavvikelsen behöver vi definiera variansen (**variance** på engelska).
- ▶ Variansen är ett mått på **variation**.

# Fördelningens spridning för numeriska variabler, forts.

- ▶ Det finns många tänkbara mått på variation.
- ▶ Ett tänkbart mått på variation är **summan av alla kvadratiska avstånd** till en central punkt  $a$ ,

$$\sum (y - a)^2 \quad (1)$$

- ▶ Varför kvadrerar vi? Antag att vi inte kvadrerar och istället använder

$$\sum (y - a)$$

som mått på variation. Om vi har två datapunkter, t.ex 1 och 7 som ligger på varsin sida om  $a = 4$ , med samma avstånd till  $a$  (3 i det här fallet), så kommer de att ta ut varandra i summan eftersom

$$1 - 4 + 7 - 4 = 0.$$

- ▶ Om vi istället kvadrerar får vi

$$(1 - 4)^2 + (7 - 4)^2 = 2 \cdot 3^2,$$

och de båda bidrar lika mycket till summan i (1) (var och en med  $3^2 = 9$ ).

# Fördelningens spridning för numeriska variabler, forts.

- ▶ Vilket central punkt  $a$  ska vi välja?
- ▶ Man kan visa att om vi väljer medelvärdet  $\bar{y}$  som  $a$ , dvs

$$\sum (y - \bar{y})^2$$

**så uppnår man den minsta möjliga kvadratsumman**, dvs det finns inget annat val av  $a$  som kan göra  $\sum (y - a)^2$  mindre.

- ▶ Denna egenskap (att minimera kvadratsumman) kallas för minsta kvadrat egenskapen (**least squares property** på engelska).
- ▶ Variansen  $s^2$  för ett stickprov definieras som (nästan) **medel kvadrataavståndet till stickprovets medelvärde** enligt

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

- ▶ Varför nästan medel och inte bara medel? Notera division med  $n - 1$  istället för  $n$  som det hade varit om det var medel.

# Fördelningens spridning för numeriska variabler, forts.

- ▶ Varför delar vi med  $n - 1$  istället för  $n$ ? Vi behöver lära oss mer statistik innan vi kan förklara det.
- ▶ Problemet med variansen är att eftersom vi kvadrerar observationer när vi räknar  $s^2$  så kommer dess enhet att vara kvadrerad.
- ▶ Svårt att tolka enheter i kvadrat.
  - ▶ Om vi mäter månadslöner så är enheten för variansen kronor i kvadrat.
  - ▶ Om vi mäter ålder så är enheten för variansen ålder i kvadrat.
- ▶ **Standardavvikelsen**  $s$  ger oss ett spridningsmått i samma enhet som data,

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}.$$

# Vad händer om vi får nya data?

- ▶ Det vi har gått igenom på denna föreläsning kallas ofta för deskriptiv statistik (**descriptive statistics** på engelska).
- ▶ Deskriptiv statistik **utgår från ett enda stickprov** och sammanfattar dess egenskaper, genom att använda allt från stapeldiagrammen till variansen (och allt där emellan vi har gått igenom), beroende på om variabeln är kategorisk eller numerisk.
- ▶ Det är viktigt att förstå att allt vi har gått igenom, från stapeldiagrammen till variansen och allt där emellan, **har grundat sig på ett enda stickprov**.
- ▶ Om vi tar ett nytt stickprov, så kan vi följa samma recept för att producera allt från stapeldiagrammen till variansen (och allt där emellan). **Resultaten kommer att variera från stickprov till stickprov**.
- ▶ Vi kommer att få se åtskilliga exempel på detta under kursens gång.

# Vad händer om vi får nya data?, forts.

- ▶ Del II av kursen behandlar inferentiell statistik (**inferential statistics** på engelska) som formellt studerar hur resultaten varierar från stickprov till stickprov genom att modellera den **teoretiska fördelningen av data**.
- ▶ Den teoretiska fördelningen av data tillåter oss att förstå hur resultaten varierar från stickprov till stickprov **utan att ha tillgång till flera stickprov**.
- ▶ Slutprodukten efter båda kurser är att ni kan använda inferentiell statistik för att dra generella slutsatser om en population med hjälp av ett endaste stickprov.



# References I

De Veaux, R. D., Velleman, P., and Bock, D. (2021). *Stats: Data and Models*. Pearson, Harlow, United Kingdom, fifth edition.