Statistik och dataanalys I

F5: Standardiserade variabler och normalfördelningen

Valentin Zulj

Statistiska institutionen

Stockholm University

Vad har vi gjort hittills?

- Hittills har vi jämfört olika numeriska fördelningar
- Vi har ställt frågor i stil med
  - Hur kan vi visa skillnader mellan olika numeriska fördelningar med **låddiagram (boxplot)**?
  - Hur bedömer vi om en skillnad mellan grupper är **slumpmässig**?
  - Hur kan vi **transformera** värden så att skillader syns tydligare i ett diagram?

Vad ska vi göra nu?

- I dagens föreläsning ska vi fortsätta gräva i jämförelser
- Vi kommer fråga oss om det går att jämföra värden som är mätta på helt olika skalor/enheter
  - Går det t.ex. att göra en meningsfull jämförelse mellan **vikten** på en båt och **längden** på en lastbil?
  - *Ja, faktiskt!* Men för det behöver vi en ny typ av måttstock
- Den nya måttstocken kallas för **Z-värde**, och är ett resultat av **standardisering**
- Z-värden och standardisering kommer leda oss till **normalfördelningen**, som är central för statistiken

Z-värden och standardisering

Z-värde - ett mått med standardavvikelser som enhet

- *Z*-**värdet** (*en: z-score*) mäter avvikelsen från genomsnittet för en variabel – mätt i antalet standardavvikelser
- Istället för att mäta i variabelns egna enheter (t.ex. kg, m, kr) använder z-värdet standardavvikelsen som enhet **hela tiden**
- Med z-värden kan vi jämföra variabler som är mätta på helt olika skalor!
- Vi börjar med att ge ett lite mer intuitivt exempel, och tittar sedan närmare på detaljer och beräkningar

Z-värde – Exempel med friidrott

- Kursboken (sid 152) jämför en länghoppare som hoppar 6.58 meter med en löpare som springer 200 meter på 23.26 sekunder, båda i OS 2016
- Vi kan inte jämföra en längd i meter med en tid i sekunder
- Men: vi **kan** jämföra hur många **standardavvikelser** respektive prestation avviker från den **genomsnittliga deltagaren** i respektive tävling
- Vi kan räkna ut att vinnaren i längdhopp hoppade **1.66 standardavvikelser längre** än det genomsnittliga hoppet i tävlingen
- Vi kan också räkna ut att vinnaren i 200 meter sprang på en tid som var **2.02 standardavvikelser mindre** än den genomsnittliga tiden i tävlingen
- **Slutsats:** Vinnaren i 200 meter gjorde en mer imponerande insats, åtminstone i relation till övriga insatser som gjordes i respektive gren
- Vi jämför fortfarande varje insats med genomsnittet i sin gren, men då vi transformerar allt till gemensam skala kan vi jämföra **mellan** grenar också!

- Z-värdet mäter som sagt avvikelser med standardavvikelse som enhet, och bygger därför på standardavvikelsen

Repetition av stadardavvikelse: För en numerisk variabel y är standardavvikelsen

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

I uttrycket ovan står s_y^2 för variansen, som vi beräknar som

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

Vi kan också beräkna allt i ett enda steg, alltså

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

Att räkna ut z-värdet

- Vi konstaterade att vinnaren i längdhopp hoppade 1.66 standardavvikelser längre än det genomsnittliga hoppet i tävlingen
- Med andra ord hade vinnarens bästa hopp ett z-värde på 1.66
- **Men:** Hur har vi räknat ut det?

Defintion av z-värde: Antag att vi har gjort n observationer av en variabel, $y_1, \ldots, y_n$, och låt $\bar{y}$ och s beteckna genomsnittet respektive standardavvikelsen för y

Vi definierar då z-värdet för den enskilda observationen y_i som

$$z_i = \frac{y_i - \bar{y}}{s}$$

- För att räkna på z-värden behöver vi en numerisk variabel/fördelning – **varför?**
- Det är inte givet hur vi skulle beräkna t.ex. standardavvikelsen för en kategorisk variabel!

Att räkna ut z-värdet

Fortsättning av friidrottsexempel: Vi har att den genomsnittliga hopplängden i tävlingen var $\bar{y} = 6.17$ meter, med standardavvikelsen $s = 0.247$ meter.

Vi använder formeln från föregående slide för att få

$$z_i = \frac{y_i - \bar{y}}{s} = \frac{6.58 - 6.17}{0.247} = 1.66.$$

Vinnaren hoppade 1.66 standardavvikelser högre än genomsnittshoppet i tävlingen.

I 200-meterstävlingen var den genomsnittliga tiden $\bar{y} = 24.58$ sekunder, med standardavvikelsen $s = 0.654$ sekunder. z-värdet för vinnaren, som sprang på 23.26 är

$$z_i = \frac{y_i - \bar{y}}{s} = \frac{23.26 - 24.58}{0.654} = -2.02.$$

Notera att z-värdet är *negativt*, då observationen är *mindre* än genomsnittet – detta är rimligt eftersom vinnaren springer på *minst/kortast* tid!

Att översätta mellan skalor

- Att översätta från en enhet (t.ex kg eller meter) till enheten standardavvikelser är inte mer komplicerat än att översätta mellan vilka två enheter som helst
- En mile är t.ex. ≈ 1.6 kilometer – om avståndet mellan två punkter är 12 km, så är avståndet i miles $12/1.6 = 7.5$ – vi har så bytt enhet från km till miles!
- På samma sätt: om avståndet mellan y och $\bar{y}$ är 5 km, och standardavvikelsen är 2 km, så blir avståndet i standardavvikelser $5/2 = 2.5$
- En viktig skillnad är att en mile alltid är ungefär 1.6 km, men storleken på en standardavvikelse varierar från fall till fall

Att räkna ut originalvärdet med hjälp av z-värdet

- Vi har sett hur vi kan räkna ut z-värdet för en observation, alltså det antal standardavvikelser som observationen skiljer sig från genomsnittet
- Vi skulle kunna vända på frågan, och t.ex. fråga hur långt vi måste hoppa för att vårt hopp ska vara **två standardavvikelser längre än genomsnittet**
- Detta kommer visa sig vara särskilt hjälpsamt när vi börjar arbeta med normalfördelningen

Vi kan ta formeln för z-värdet, och skriva om den m.h.a. algebra, för att få

$$z = \frac{y - \bar{y}}{s} \iff y = \bar{y} + zs$$

Detta är en "anti-transformation", vi går från z-skalan tillbaka till originalskalan!

Att räkna ut originalvärdet med hjälp av z-värdet

- I längdhoppstävlingen har vi att $\bar{y} = 6.17$ meter, $s = 0.247$ meter. Hur långt behöver vi hoppa för att hamna två standardavvikelser genom snittet?
- Med andra ord, hur stort är y om $z = 2$?
- Vi använder formeln från föregående slide och får

$$y = \bar{y} + zs = 6.17 + 2 \cdot 0.247 = 6.664.$$

- Vi behöver hoppa ca 6.66 (sammanträffande) meter!

- **Fråga:** Hur långt måste du hoppa för att z-värdet ska vara större än 0?

Standardiserade variabler

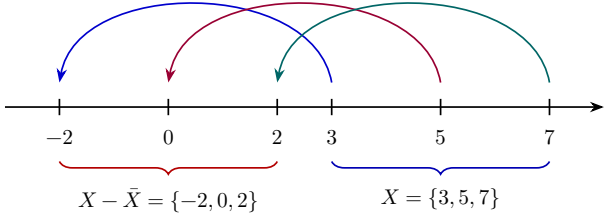
- Vi har sett att z-värdet räknas ut med formeln

$$z = \frac{y - \bar{y}}{s}$$

- Om vi har en variabel y med medelvärde $\bar{y}$ och standardavvikelse s, kan vi visa att
  - Den nya variabeln $y - \bar{y}$ har medelvärde 0, men att standardavvikelsen är fortfarande s
  - Den nya variabeln $z = (y - \bar{y})/s$ har medelvärde 0 och standardavvikelse 1
- Eftersom z alltid har medelvärde 0 och standardavvikelse 1, oavsett vilka värden vi har på variabeln y, säger vi att z är en **standardiserad variabel**
- Standardisering är en ny typ av **transformation**, som också är mycket vanlig i statistiken

Intuition bakom centrering

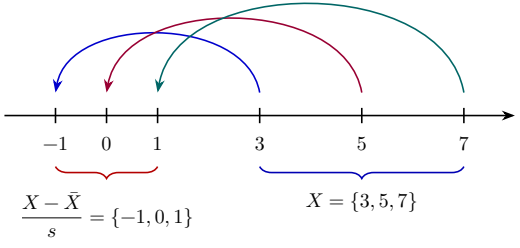
- Antag att vi har $x_1 = 3$, $x_2 = 5$, och $x_3 = 7$, vilket ger $\bar{x} = 5$ och $s = 2$
- När vi tar $x_i - \bar{x}$ flyttar vi varje enskild observation längs x-axeln, åt vänster om $\bar{x}$ är positivt och åt höger om $\bar{x}$ är negativt
- Varje x_i kommers flytta lika långt och åt samma håll



- Vi ser i figuren att $x_i - \bar{x}$ är centrerad runt 0, men att spridningen för $x_i - \bar{x}$ är vara lika stor som spridningen för x
- $x_i - \bar{x}$ har medelvärde 0, men samma standardavvikelse (s) som x själv

Intuition bakom standardisering

- Standardiseringen $(x_i - \bar{x})/s$ gör då två saker:
  1. Täljaren $x_i - \bar{x}$ flyttar värden längs x-axeln, som tidigare
  2. Division med s påverkar hur långt varje punkt flyttas – den påverkar alltså **skalan**
- Punkterna flyttas fortfarande åt samma håll, men p.g.a. steg 2 blir hoppen olika långa



- Vi ser i bilden att $(x_i - \bar{x})/s$
  1. Är centrerad runt 0 (har medelvärde 0)
  2. Har annan spridning är x själv (den nya spridningen blir automatisk $s = 1$)

Större exempel i R

- Vi kan använda R för att testa centrera/standardisera ett riktigt datamaterial
- I R finns datamaterialet **mtcars** som innehåller data för 32 bilmodeller
- Datamaterialet innehåller variabeln **mpg** som mäter bränsleförbrukning (miles per gallon)

```
> data(mtcars)                          # laddar data
> y <- mtcars$mpg                       # sparar mpg i y
>
> favstats(y) |> round(3) # beskrivande statistik, avrundad till 3 decimaler
>   min     Q1 median   Q3  max   mean    sd  n missing
>  10.4 15.425   19.2 22.8 33.9 20.091 6.027 32       0
```

- Vi ser att $\bar{y} = 20.091$ och $s = 6.027$

- Vi kan titta på den centererade versionen av y

```
> m <- mean(y)     # Sparar medelvärdet
> s <- sd(y)       # Sparar standardavvikelsen
> y_cent <- y - m # Centrerar
>
> favstats(y_cent) |> round(3) # Beskrivande statistik
>     min     Q1 median    Q3    max mean    sd  n missing
>  -9.691 -4.666 -0.891 2.709 13.809    0 6.027 32       0
```

- Och på liknande sätt kan vi titta på den standardiserade versionen

```
> z <- (y - m) / s         # Standardiserar
>
> favstats(z) |> round(3) # Beskrivande statistik
>     min     Q1 median   Q3   max mean sd  n missing
>  -1.608 -0.774 -0.148 0.45 2.291    0  1 32       0
```

- **Precis vad vi förväntade oss!**

z är en standardiserad variabel

- Vi kan nu skriva ut (de avrundade) värdena på den standardiserade variabeln

```
> z |> round(3)
>  [1]  0.151  0.151  0.450  0.217 -0.231 -0.330 -0.961  0.715  0.450 -0.148
> [11] -0.380 -0.612 -0.463 -0.811 -1.608 -1.608 -0.894  2.042  1.711  2.291
> [21]  0.234 -0.762 -0.811 -1.127 -0.148  1.196  0.980  1.711 -0.712 -0.065
> [31] -0.845  0.217
```

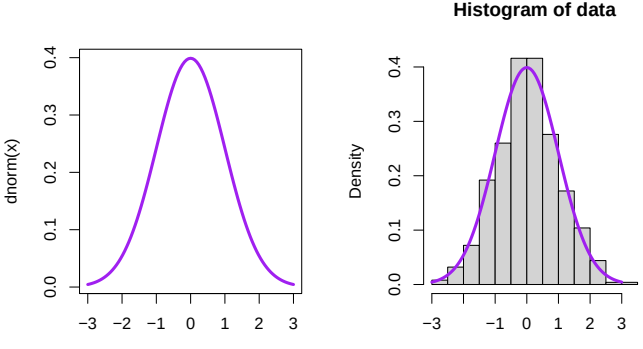
- Vi kommer ihåg att z är en standardiserad version av miles per gallon, alltså hur många miles respektive bil kan köra på en gallon av bränsle
- **Fråga:** Hur tolkar vi det första värdet ovan, d.v.s. 0.151 (hör till en Mazda RX-4)?

z är en standardiserad variabel

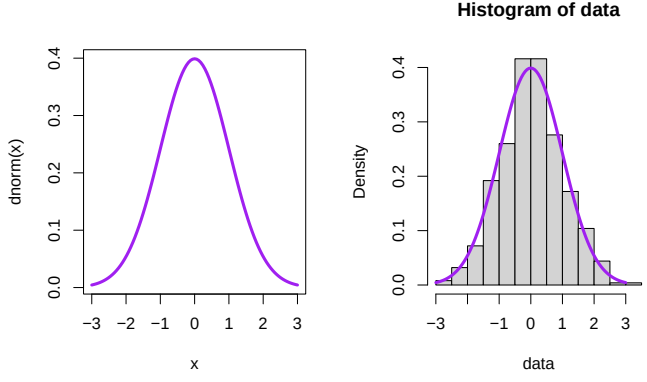
- Hittils har vi lärt oss att beräkna z-värdet
- Vi har också lärt oss att göra det omvända, dvs att räkna ut hur stort värdet på en observation måste vara för att motsvara ett visst z-värde
- Vi har sett att transformation till av y till z ger en ny variabel med medelvärde 0 och standardavvikelse 1
- Vi har lärt oss att tolka ett z-värde, och vi vet t.ex. att ett längdhopp vars z-värde är större än 0 är längre än genomsnittshoppet i tävlingen
- Men **hur** exceptionellt är det att göra ett hopp som är 2.02 standardavvikelser över genomsnittet? Hur vanligt/ovaligt är det?
- Liknande frågor är ofta av intresse, både i vardagen och i industrin
- För att svara på denna typ av fråga kan vi ta hjälp av **normalfördelningen**

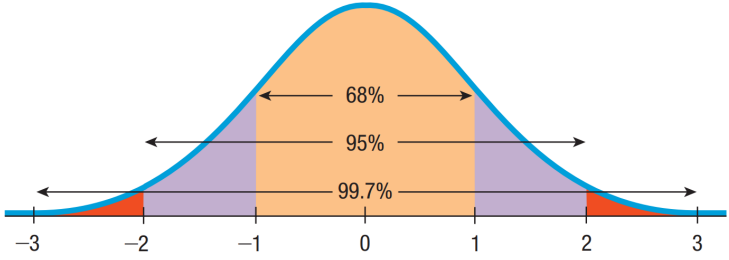
Normalfördelningen

- Under F2 tittade vi på histogram, och sa att formen på ett histogram beskriver hur värden på en variabel **fördelar sig**
- Normalfördelningen beskriver en speciell typ av fördelning
- Figuren till vänster visar en **normalfördelningskurva**, och figuren till höger är ett histogram som visar en normalfördelad variabel
- Histogrammet har ungefär samma form som normalfördelningskurvan

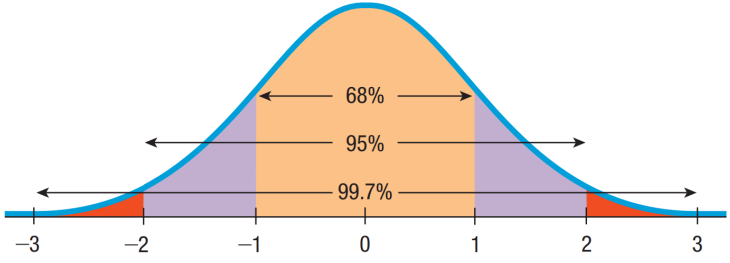


- Formen på en normalfördelningskurva påminner om en kyrkklocka, och denna typ av kurva kallas ibland också för just *bell curve*
- Fördelningen kan ibland kallas för en *Gaussisk fördelning*, men i den här kursen kommer vi konsekvent att använda termen normalfördelning



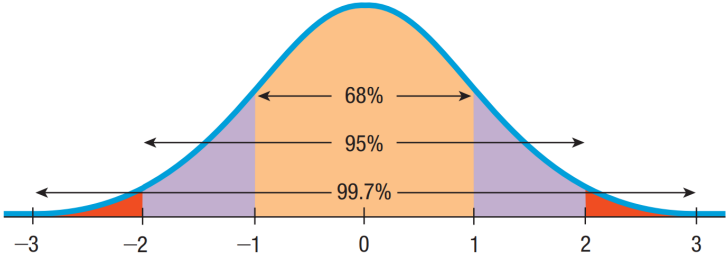


- Om en variabel är normalfördelad kan vi med hjälp av detta räkna ut hur stor andel av observationerna som ligger inom ett visst intervall
- Arean under normalfördelningskurvan representerar 100% av våra observationer
- Skalan på x-axeln i figuren visar antalet standardavvikelser från medelvärdet (z-värdet)

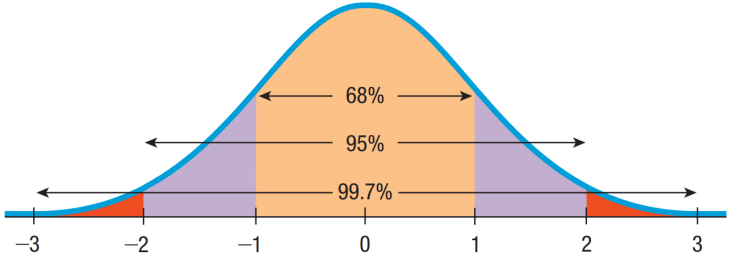


- I normalfördelningen ligger
  - $\approx 68\%$ av alla observationer inom en standardavvikelse från medelvärdet
  - $\approx 95\%$ av alla observationer inom 2 standardavvikelser från medelvärdet
  - $\approx 99.7\%$ av alla observationer inom 3 standardavvikelser från medelvärdet

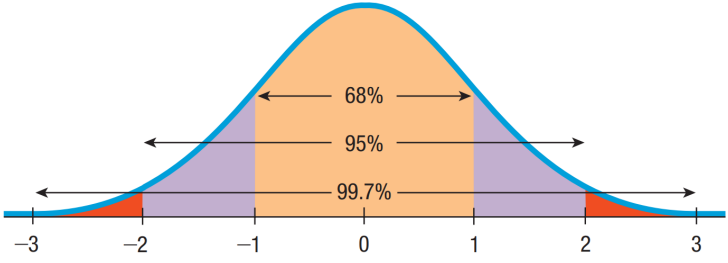
Räkneexempel 1



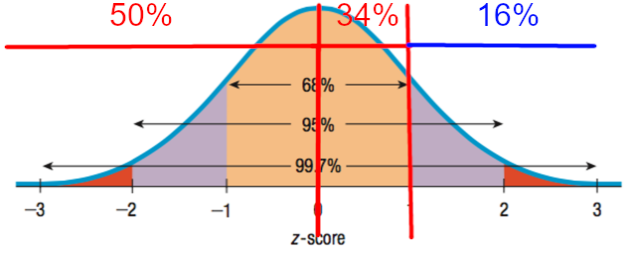
- Låt oss använda det vi vet om normalfördelningen för att göra beräkningar!
- **Antag** att hopplängderna i en längshoppstävling **är normalfördelade**
- Hur stor andel av hoppen är antingen minst två standardavvikelser större eller minst två standardavvikelser mindre än genomsnittet?



- Hur stor andel av hoppen är antingen minst två standardavvikelser större eller minst två standardavvikelser mindre än genomsnittet?
- Från figuren: om hopplängderna är normalfördelade så är 95% av hoppen i intervallet mellan -2 och 2 standardavvikelser från genomsnittet
- Andelen hopp utanför detta intervall är alltså $100\% - 95\% = 5\%$



- Hur stor andel av hoppen är minst en standardavvikelse längre än snittet?
- För att lösa denna uppgift behöver vi introducera en ny egenskap hos normalfördelningen, nämligen att den är **symmetrisk**
- Vi ser att normalfördelningskurvan speglar sig runt värdet 0 på x-axeln, det är alltså lika många observationer till vänster om 0 som till höger om 0
- Vi har så att 50% av observationerna är mindre än 0, och 50% är större än 0



- Med hjälp av fördelningens symmetri kan vi komma fram till att
  - Av hoppen som ligger mellan -1 och 1, ligger hälften till höger om 0 (alltså $68\%/2 = 34\%$)
  - 50% av alla hopp ligger till höger om noll, och 34% ligger mellan 0 och 1
  - Vi får därför att andelen hopp till höger om 1 är $50\% - 34\% = 16\%$
- Andelen hopp som är minst en standardavvikelse längre än snittet är så 16%

- Vi har så omvandlat den relativt vaga utsagan “hoppet är minst en standardavvikelse längre än snittet”, till det mer konkreta “hoppet är bland de 16% bästa”
- Vi har gjort detta genom att anta att **hopplängderna är normalfördelade**
- Om hopplängderna **inte** är normalfördelade kan vi **inte** konkretisera på samma sätt!
- Om vi har data och ser i ett histogram att deras fördelning inte liknar normalfördelningen så blir det **direkt felaktigt** att använda den typ av beräkningar som gjorts ovan
- Felen blir större och större ju “längre bort” från normalfördelningen vi kommer, och kan t.ex. bli väldigt stora om fördelningen är ordentligt skev

Räkneexempel 3

- Vi antar att genomsnittshoppet är $\bar{y} = 4.2$ meter, och att $s = 0.4$ meter
- Hur långt måste ett hopp vara **i meter** för att vara bland de 16% längsta hoppen?
- Det här är ett mer intressant exemepel än de tidigare: frågan är ställd i vanliga enheter (alltså meter) och inte standardavvikelser
- Detta är intressant då vi vanligtvis tänker i vanliga enheter, men vi har bara pratat om normalfördelningen i enheten standardavvikelser
- Vad kan vi göra för att lösa detta problem?
- Vi kan översätta mellan skalor!

- Hur långt måste ett hopp vara **i meter** för att vara bland de 16% längsta hoppen, när $\bar{y} = 4.2$ meter, och $s = 0.4$ meter?
- För att vara bland de 16% längsta hoppen måste hoppet vara minst 1 standardavvikelse längre än genomsnittet (från Räkneexempel 2)
- Vi vill översätta gränsen till meter, och använder formeln från innan

$$z = \frac{y - \bar{y}}{s} \iff y = \bar{y} + zs$$

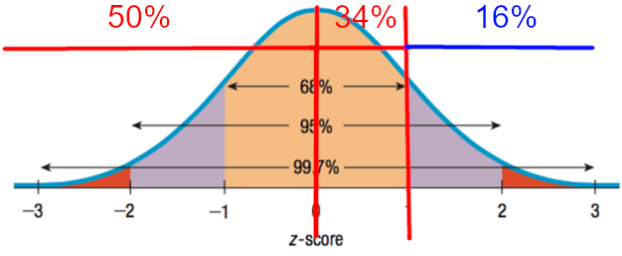
- Hoppet måste alltså vara $\bar{y} + 1 \cdot s$ meter för att vara en standardavvikelse från snittet
- Vi stoppar in våra siffror, och får gränsen $\bar{y} + 1 \cdot s = 4.2 + 1 \cdot 0.4 = 4.6$ meter!

Percentiler

- I F2 pratade vi lite kort om **percentiler**
- Om vi har en numerisk variabel så är en percentil ett värde som är större än en viss specificerad procentandel av observationerna

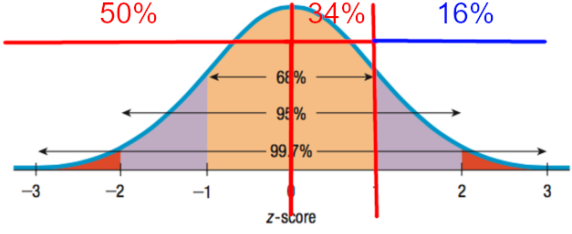
- Den 75:e percentilen är ett värde som är större än ungefär 75 procent av observationerna och mindre än ungefär 25 procent av observationerna
- Den 75:e percentilen är alltså samma sak som den tredje kvartilen Q_3

Exempel med percentil i Normalfördelningen



- I vår givna tävling måste ett längdhopp vara minst 4.6 meter för att ligga precis **en standardavvikelse över genomsnittet**
- Från bilden: ett hopp som är precis en standardavvikelse över snittet är även
  - Längre än ungefär 84% av hoppen i tävlingen
  - Kortare än ungefär 16% av hoppen
- **Fråga:** Hur kan vi uttrycka detta i termer av percentiler?
- Ett hopp på 4.6 meter ligger vid den **84:e percentilen**, givet att hoppen är normalfördelade

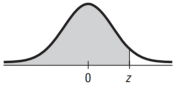
Percentiler i normalfördelningen



- Antag att vi vill veta hur långt ett hopp måste vara för att vara topp 10% i tävlingen, dvs ligga över den 90:e percentilen
- Bilden visar att hoppet
  - Måste vara mer än 1 standardavvikelse över snittet (84:e percentilen)
  - Inte behöver vara så långt som 2 standardavvikelser över snittet (97.5:e percentilen)
- För ett mer exakt svar behöver vi en **normalfördelningstabell**

- Normalfördelningstabellen (utklipp nedan) kan användas för att **översätta z-värden till proportioner, eller proportioner till z-värden**
- Vi får ett z-värde genom att kombinera talen i den vänstra marginalen med talen i den övre marginalen
- För varje z-värde anger tabellen den andel som **ligger till vänster** om detta z-värde

Table Z (cont.)
Areas under the standard Normal curve



	Second decimal place in z									
z	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

Att läsa normalfördelningstabeller

				Second dec	
z	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700

- Vi tittar på den markerade cellen i tabellen:
  - Den står på raden med 0.3 i vänstermarginalen
  - Den finns i kolumnen där med 0.02 i den övre marginalen
- 0.3 och 0.02 sätts samman till 0.32 – talet i den övre marginalen anger den andra decimalen i det z-värde som den markerade cellen avser

Att läsa normalfördelningstabeller

				Second dec	
z	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700

- Det gulmarkerade talet är 0.6255, och betyder att z-värdet 0.32 har 62.55% av alla observationer till **vänster** om sig i normalfördelningen
- Vi kan uttrycka det som att z-värdet 0.32 ligger vid den 62.55:e percentilen

				Second dec	
z	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700

- Vi behåller kopplingen till vår längdhoppstävling, och tolkar tabellvärdet i det sammanhanget
- Om någon gör ett hopp som är 0.32 standardavvikelser längre än snittet, så är det hoppet längre än 62.55% av alla hopp i tävlingen

- Vi vill sällan åt z-värdet i sig, utan det är bara ett led i våra beräkningar
- Z-värdet låter oss översätta mellan originalenheten på vår variabel (m, kg, kr, etc) och en andel i procent, vilket kan vara hjälpsamt
- Om vi har ett värde i originalskalan och vill veta andelen observationer som är mindre eller större än detta värde kan vi använda följande relationer

$$\text{y-värde} \implies \text{z-värde} \implies \text{andel i procent}$$

- Om vi vill veta vad värdet i originalskalan behöver vara för att en bestämd andel av observationerna ska vara mindre/större så kan vi gå bakvägen

$$\text{andel i procent} \implies \text{z-värde} \implies \text{y-värde}$$

- Låt oss göra en uträkning av varje slag!

I längdhoppstävlingen har vi $\bar{y} = 4.2$ meter och $s = 0.4$. Antag att vi gör ett hopp om $y_i = 4.5$ meter, och att vi söker andelen hopp som är kortare respektive längre än vårt.

Vi har ett värde på originalskala och söker en andel, alltså jobbar vi enligt

$$\text{y-värde} \implies \text{z-värde} \implies \text{andel i procent}$$

Vi börjar med att beräkna z-värdet med den vanliga formeln

$$z_i = \frac{y_i - \bar{y}}{s} = \frac{4.5 - 4.2}{0.4} = 0.75.$$

Vårt hopp var alltså 0.75 standardavvikelser längre än genomsnittshoppet

Först vill vi hitta andelen hopp som har ett **mindre** z-värde än 0.75, och använder normalfördelningstabellen rakt av (den ger andelar *till vänster* om ett vistt värde)

Vi tittar på raden med 0.7 i vänstermarginalen, och kolumnen med 0.05 i övre marginalen, och hittar 0.7734

Ungefär 77.34 procent av alla hopp i tävlingen är kortare än vårt hopp, vilket innebär att det ligger ungefär vid den 77:e percentilen i förhållande till övriga hopp i tävlingen

Vi vet att 77.34% av alla hopp har mindre z-värde än 0.75, vilket medför att $100 - 77.34 = 22.66\%$ måste ha ett z-värde som är större än 0.75

Alltså: 22.66% av alla hopp är längre än 4.5 meter!

Normalfördelningstabeller – räkneexempel 2

- Hur långt måste vi hoppa för att hamna bland de 10% bästa?
- Vi har nu en andel och söker ett värde på originalskala, och räknar därför enligt

$$\text{andel i procent} \implies \text{z-värde} \implies \text{y-värde}$$

- Vi vill hoppa så långt att minst 90 procent av hoppen i tävlingen är kortare än vårt hopp, alltså vill vi ligga över (till höger om) den 90:e percentilen
- Vi letar efter andelen 0.9 i normalfördelningstabellen (exakt 0.9 finns inte i tabellen, men vi tar 0.8997 som är närmast, på rad *1.2* och kolumn *0.08*)
- $z = 1.28$ betyder att hoppet måste vara minst 1.28 standardavvikelser längre än genomsnittet för att vara topp 10%
- Nu återstår bara att omvandla z-värdet till meter, med formeln

$$y_i = \bar{y} + zs = 4.2 + 1.28 \cdot 0.4 = 4.712$$

- **Slutsats:** Vi måste hoppa längre än 4.712 meter för att vårt hopp ska vara topp 10%

- I `R` är funktionerna **qnorm** och **pnorm** kopplade till normalfördelningstabeller
- Vi använder `pnorm()` när vi har ett z-värde och vill veta hur stor andel av observationerna i en normalfördelning som har ett mindre z-värde
- I ett av våra exempel ville vi veta hur stor andel av längdhoppen som hade ett mindre z-värde än 0.75, och tabellen gav oss då 77.34 procent
- I `R` gör vi motsvarande beräkning på följande sätt

```
> pnorm(0.75)
> [1] 0.7733726
```

Normalfördelningstabeller i R

- I `R` är funktionerna **qnorm** och **pnorm** kopplade till normalfördelningstabeller
- Vi använder `qnorm()` för att se vilket z-värde som är större än en bestämd andel av observationerna i en normalfördelning
- Tidigare såg vi att 1.28 är det z-värde som är större än 90% av observationerna (och mindre än övriga 10%)
- I `R` gör vi motsvarande beräkning på följande sätt (notera att vi i R skriver 0.9 istället för 90%)

```
> qnorm(0.9)
> [1] 1.281552
```

Varför just normalfördelningen?

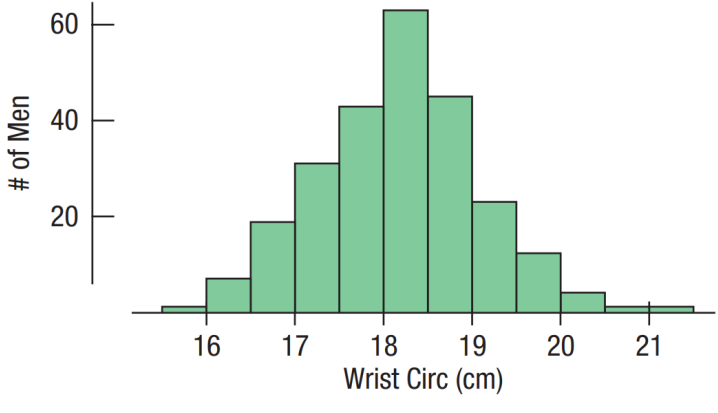
- En numerisk variabel kan vara fördelad på många olika sätt
- Varför är vi så intresserade av just variabler som är normalfördelade?
- Den franske matematikern Abraham de Moivre visade på 1700-talet att många fördelningar i verkligheten är lika normalfördelningen (som då hette något annat)
- Dessutom är det så att **medelvärden** ofta fördelar sig enligt en normalfördelning
- Det förhållandet kallas **centrala gränsvärdessatsen (the Central Limit Theorem)** och kommer att vara viktigt i *del 2* av kursen

Undersök om en variabel är normalfördelad

- Normalfördelningen är en förutsättning för beräkningarna i våra räkneexempel, men vi kan inte utan argument utgå från att en numerisk variabel är normalfördelad
- När vi använder normalfördelningen för våra beräkningar måste vi alltså först undersöka om vår variabel verkligen är normalfördelad
- Vi ska nu gå igenom några sätt att undersöka om en variabel är normalfördelad, eller åtminstone om den följer en fördelning som liknar en normalfördelning

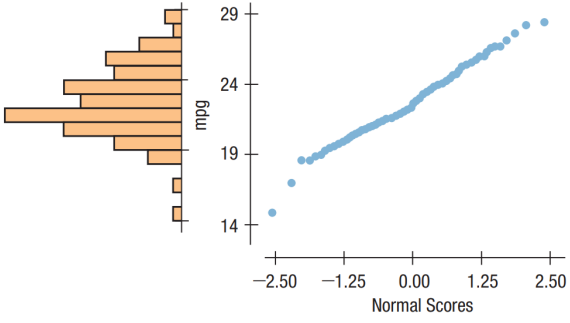
Undersök om en variabel är normalfördelad

- Vi kan säga att en fördelning är **nästan normalfördelad** (*nearly normal*) om den är symmetrisk, bara har en topp (unimodal) och inte har tydliga outliers
- Fördelningen på bilden får sägas leva upp till villkoren för att vara nästan normalfördelad, men det är till viss del en subjektiv bedömning



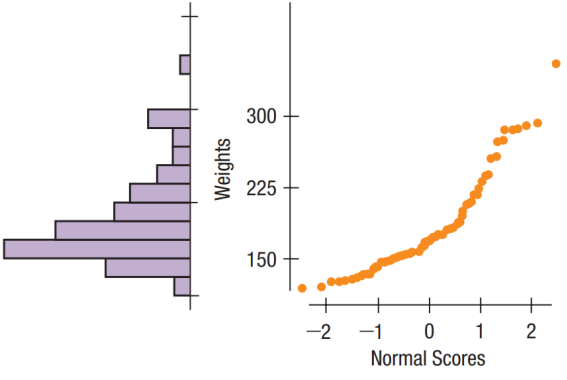
Undersök om en variabel är normalfördelad

- Vi kan också använda en **normalfördelningsplot** (*normal probability plot*) – om variabeln är normalfördelad bör punkterna i plotten ligga längs en rät linje
- **Exempel:** Bensinförbrukning i mpg (miles per gallon) för en Nissan Maxima, insamlad under 8 år av en av kursbokens författare
- Linjen är någorlunda rak – två observationer till vänster är mindre än vad de borde vara, men det är ändå rimligt att se variabeln som normalfördelad



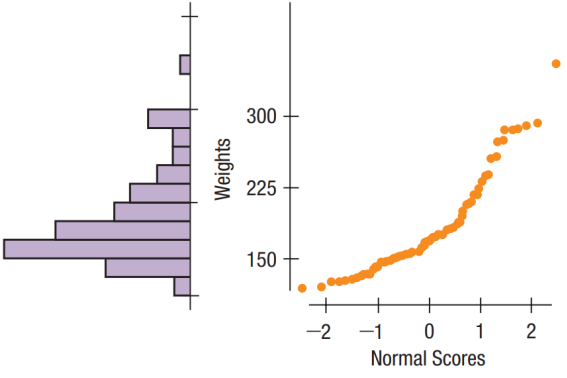
- **Kom ihåg!**
  - Även om en variabel är ungefär normalfördelad så följer den förmodligen inte exakt en normalfördelning
  - Var medveten om att resultaten av beräkningarna därför inte kan förväntas vara så exakta, men förhoppningsvis ge en bra approximation

Undersök om en variabel är normalfördelad



- Ovan ser vi en normalfördelningsplot och ett histogram ur De Veaux et al. (2021)
- Linjen är inte rak, och dessutom ser vi i histogrammet att fördelningen inte är symmetrisk, utan skev åt höger
- Det vore orimligt att betrakta denna variabel som normalfördelad, och om vi räknade med den som normalfördelad skulle resultaten bli missvisande

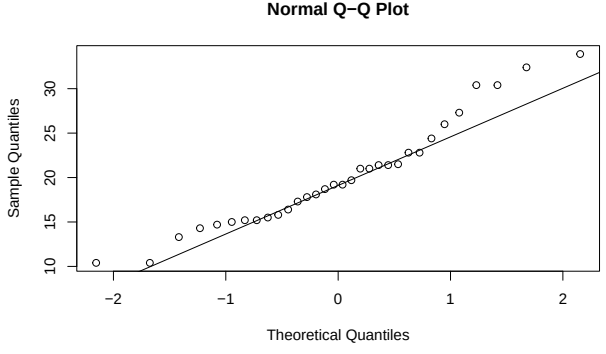
Undersök om en variabel är normalfördelad



- Det kan gå att transfromera för att få en variabel som mer liknar något normalfördelat
- Vi har tidigare sett att transformationer kan användas för att underlätta jämförelser mellan variabler med låddiagram
- Vi kan också transformera för att komma närmare en normalfördelning!

- I R kan vi använda **qqnorm** och **qqline** tillsammans för att göra en normalfördelningsplot, med en linje som visar hur punkterna bör ligga

```
> data(mtcars)
> qqnorm(mtcars$mpg)
> qqline(mtcars$mpg)
```



Tack för idag!!