Lecture 5

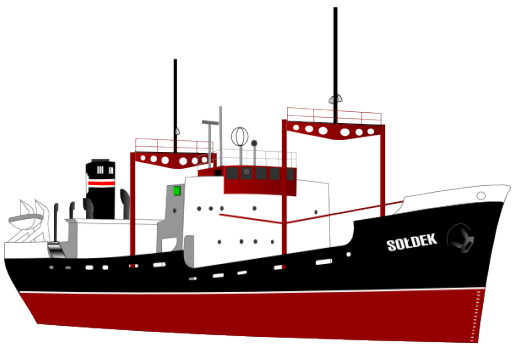
karl.sigfrid@stat.su.se

Vad har vi gjort hittills, och vad vi ska göra nu

- Hittills har vi jämfört olika numeriska fördelningar, och ställt frågor som:
  - 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 har vi gjort hittills, och vad vi ska göra nu

- Är det möjligt att också jämföra värden som har helt **olika enheter**?
- Kan vi 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.



$>$

?

z-värde - ett mått med standardavvikelser som enhet

Z-värdet (z-score) mäter avvikelsen från genomsnittet för en variabel mätt i antalet standardavvikelser.

Exempel

- Kursboken (sid 152) jämför en länghoppare som hoppar 6.58 meter med en löpare som springer 200 meter på 23.26 sekunder i OS 2016.
- Vi kan inte jämföra en längd i meter med en tid i sekunder.
- Däremot kan vi jämföra hur många **standardavvikelser** som respektive prestation avviker från **genomsnittet** bland deltagarna i respektive tävling.

z-värde - ett mått med standardavvikelser som enhet

Exempel, fortsättning

- 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 metersfinalen sprang på en tid som var **2.02 standardavvikelser mindre** än den genomsnittliga tiden i tävlingen.
- **Slutsats:** Vinnaren i 200-metesfinalen gjorde en mer imponerande insats, åtminstone i relation till övriga insatser som gjordes i de båda tävlingsgrenarna.

Standardavvikelser - att räkna ut standardavvikelsen

På föreläsning 2 tittade vi på formeln för att räkna ut standardavvikelsen.

För en numerisk variabel y är standardavvikelsen

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

där s_y^2 är variansen. s_y^2 räknas ut

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

Om vi hellre vill räkna ut standardavvikelsen i ett enda steg blir formeln

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

Att räkna ut z-värdet

- Vi konstaterade att vinnaren i längdhopp hoppade 1.66 standardavvikelse längre än det genomsnittliga hoppet i tävlingen. Längdhopparens bästa hopp hade alltså ett z-värde på 1.66. *Hur räknade vi ut det?*
- Z-värdet för en observation (exempelvis ett hopp i längdhopp) kan räknas ut **om vi känner till genomsnittet och standardavvikelsen** i en numerisk fördelning. Om värdet på vår observation betecknas y, då är z-värdet

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

där $\bar{y}$ är genomsnittet för variabeln y, och s är standardavvikelsen.

Att räkna ut z-värdet

- I vårt exemepel var den genomsnittliga hopplängden var $\bar{y} = 6.17$ meter och standardavvikelsen var $s = 0.247$ meter.
- Z-värdet för ett hopp på $y = 6.58$ meter blir då

$$z = \frac{y - \bar{y}}{s} = \frac{6.58 - 6.17}{0.247} = 1.66$$

Att räkna ut z-värdet

- Vi gör samma beräkning för vinnaren i 200-meterslöpning för att bekräfta att z-värdet för den bästa löptiden är 2.02.
- Den vinnande tiden var 23.26 sekunder. Den genomsnittliga tiden var $\bar{y} = 24.58$ sekunder och standardavvikelsen var $s = 0.654$ sekunder. Därför blir z-värdet för segertiden:

$$z = \frac{y - \bar{y}}{s} = \frac{23.26 - 24.58}{0.654} = -2.02$$

Notera att z-värdet vi räknade ut är **negativt**. Det betyder att observationen som vi räknade ut z-värdet för är **mindre** än genomsnittet. När de handlar om tiden för ett lopp är 2.02 standardavvikelser mindre samma sak som 2.02 standardavvikelser bättre.

Att räkna ut z-värdet

- Att översätta från en annan enhet, exempelvis kg eller meter, till enheten standardavvikelser fungerar på samma sätt som att översätta mellan vilka två enheter som helst.
- En amerikansk mile är exempelvis omkring 1.6 kilometer. Om avståndet mellan två punkter är 12 km och varje mile är 1.6 km så är avståndet i miles $\frac{12}{1.6} = 7.5$
- På samma sätt: om avståndet mellan y och $\bar{y}$ är 5 km och varje standardavvikelse är 2 km så blir avståndet i standardavvikelser $\frac{5}{2} = 2.5$.
- En skillnad är att en mile alltid är ungefär 1.6 km. Storleken på en standardavvikelse skiljer sig från fall till fall.

Att räkna ut z-värdet

- Vi har sett hur vi kan räkna ut z-värdet för en observation, alltså det antalet standardavvikelser som observationen skiljer sig sig från genomsnittet.
- Vi kan också vilja besvara frågor som: hur långt måste du hoppa i längd för att hoppa **två standardavvikelser över genomsnittet**?
- Det kan vi se genom att skriva om vår formel:

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

Om det genomsnittliga hoppet är $\bar{y} = 6.17$ meter och standardavvikelsen är $s = 0.247$ måste du alltså hoppa $6.17 + 2 \cdot 0.247 = 6.664$ meter för att ditt hopp ska vara två standardavvikelser över genomsnittet.

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

z är en standardiserad variabel

Vi har sett att z-värdet räknas ut med formeln

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

- Anta att vi har en variabel y med medelvärdet $\bar{y}$ och standardavvikelsen s.
- Variabeln $y - \bar{y}$ har då medelvärdet 0, och variabelns standardavvikelse är fortfarande s.
- Variabeln $z = (y - \bar{y})/s$ har medelvärdet 0 och standardavvikelsen 1.
- På grund av att z har medelvärdet 0 och standardavvikelsen 1, oavsett vilka värden vi har på variabeln y, säger vi att z är en **standardiserad variabel**.

z är en standardiserad variabel

Vi visar i R att det vi sagt om medelvärdet och standardavvikelsen för z stämmer. Vi läser in ett dataset som innehåller 32 bilmodeller. Den första bilmodellen är exempelvis en Mazda RX4. För varje modell skiver vi ut variabeln *mpg* (miles per gallon).

```
suppressWarnings(library(mosaic))
data(mtcars)
y <- mtcars$mpg
y
```

```
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8
[12] 16.4 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5
[23] 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4
```

z är en standardiserad variabel

Vi tittar på några mått som sammanfattar vår variabel:

```
favstats(y) |> round(3) #Vi avrundar 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 variabels medelvärde är drygt 20 mpg, och standardavvikelsen är ungefär 6 mpg.

z är en standardiserad variabel

- För varje observation av y subtraherar vi vi medelvärdet och delar med standardavvikelsen.
- Vi skriver ut våra värden för den nya variabeln z.

Tolka: Vad säger dessa värden exempelvis om modellen Mazda RX4, som är den första bilen i vårt dataset?

```
z <- (y - mean(y)) / sd(y)
z |> round(3)
```

```
 [1]  0.151  0.151  0.450  0.217 -0.231 -0.330 -0.961  0.715
 [9]  0.450 -0.148 -0.380 -0.612 -0.463 -0.811 -1.608 -1.608
[17] -0.894  2.042  1.711  2.291  0.234 -0.762 -0.811 -1.127
[25] -0.148  1.196  0.980  1.711 -0.712 -0.065 -0.845  0.217
```

z är en standardiserad variabel

Vi tittar på samma mått som på förra bilden, men nu för z.

```
favstats(z) |> round(3) #Vi avrundar till 3 decimaler
```

```
    min     Q1 median   Q3   max mean sd  n missing
 -1.608 -0.774 -0.148 0.45 2.291    0  1 32       0
```

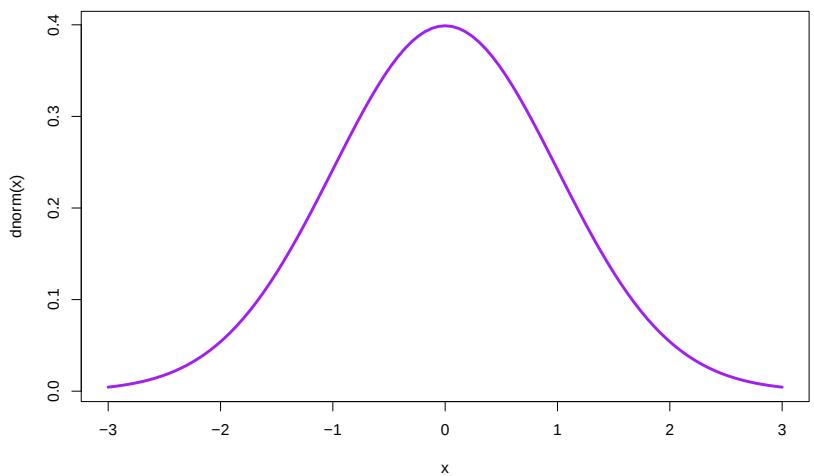
Vi har nu medelvärdet 0 och standardavvikelsen 1. Det överensstämmer med vad vi kunde förvänta oss utifrån teorin.

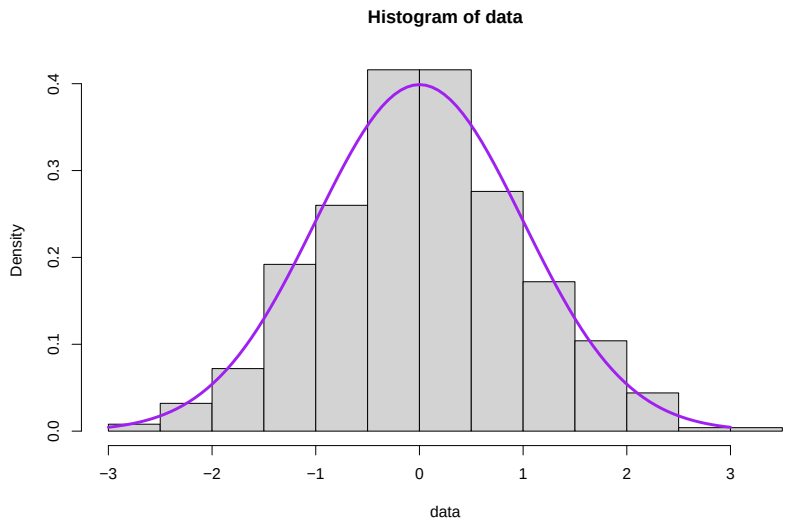
z är en standardiserad variabel

- Hittils har vi lärt oss att räkna ut z-värdet.
- Vi har också lärt oss att göra det omvända, dvs att räkna ut hur stort värdet på en variabel måste vara för att motsvara ett visst z-värde.
- Vi sett att om vi omvandlar alla värden i en variabel y till z-värden så får vi en ny variabel med medelvärdet 0 och standardavvikelsen 1.
- Däremot har vi inte sagt så mycket om **betydelsen** av z-värdet. Vi vet att om en längdhoppare gör ett hopp vars z-värde är större än 0 har lyckats bättre än genomsnittet, men **hur** bra är det att göra ett hopp som är 2.02 standardavvikelser över genomsnittet?
- För att svara på den frågan ska vi ta hjälp av **normalfördelningen**.

Normalfördelningen

- På föreläsning 1 tittade vi på histogram. Formen på ett histogram beskriver hur värden på en variabel **fördelar sig**.
- Figuren till vänster är en **normalfördelningskurva**. Figuren till höger är ett histogram som visar en normalfördelad variabel. Vi ser att histogrammet har ungefär samma form som normalfördelningskurvan.

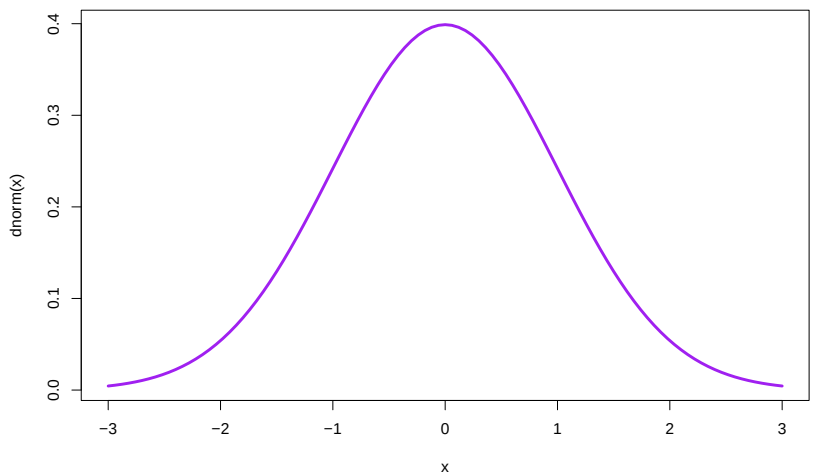


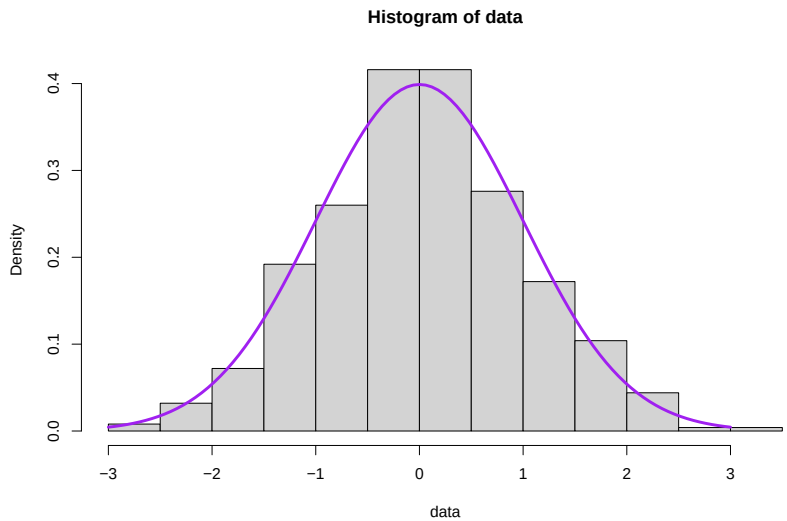


Stockholm University

Normalfördelningen

Formen på en normalfördelning påminner om en kyrkklocka, så formen och kallas ibland *bell curve*. Fördelningen kan ibland kallas för en *Gaussisk fördelning*, men i den här kursen kommer vi konsekvent att använda termen normalfördelning.

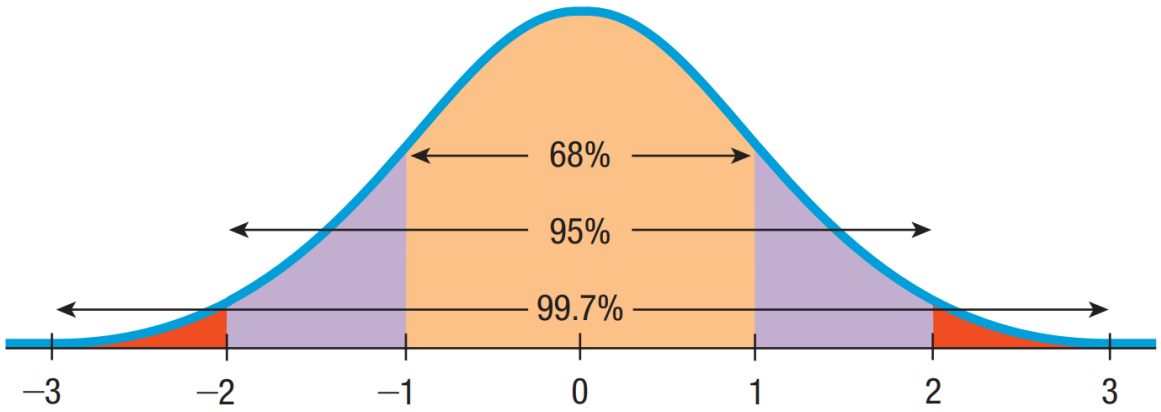




Stockholm University

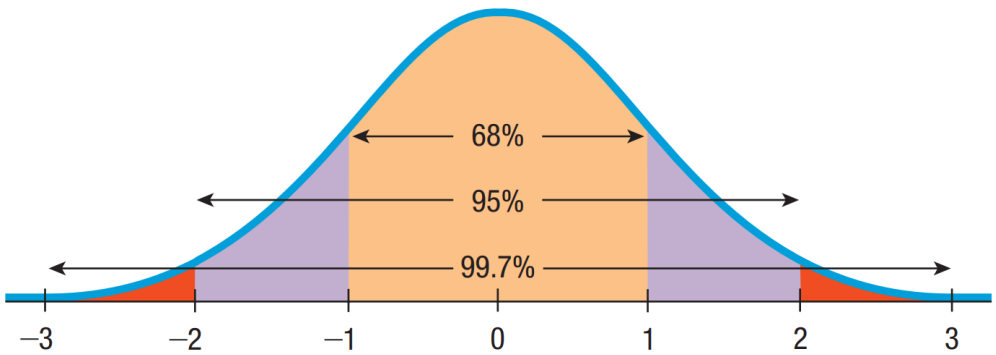
Normalfördelningen

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



Normalfördelningen

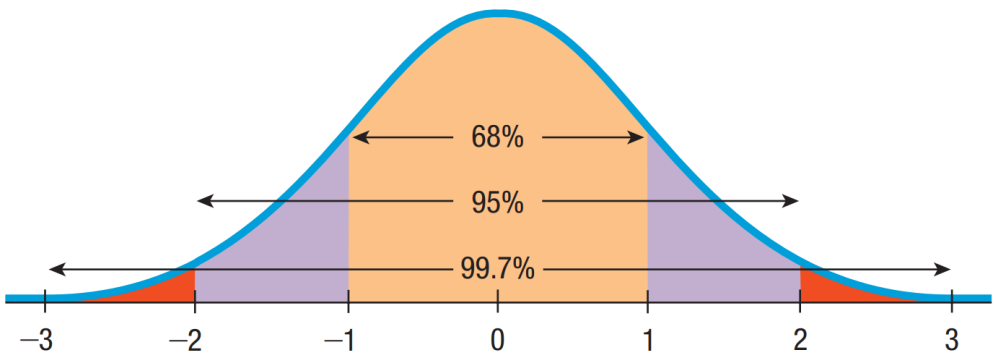
- Figuren visar att
  - 68% av alla observationer ligger inom en standardavvikelse från medelvärdet.
  - 95% av alla observationer ligger inom 2 standardavvikelser från medelvärdet.
  - 99.7% av alla observationer ligger inom 3 standardavvikelser från medelvärdet.



Stockholm University

Normalfördelningen - räkneexempel

- Låt oss använda det vi vet om normalfördelningen för att **göra beräkningar!**
- Anta 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?

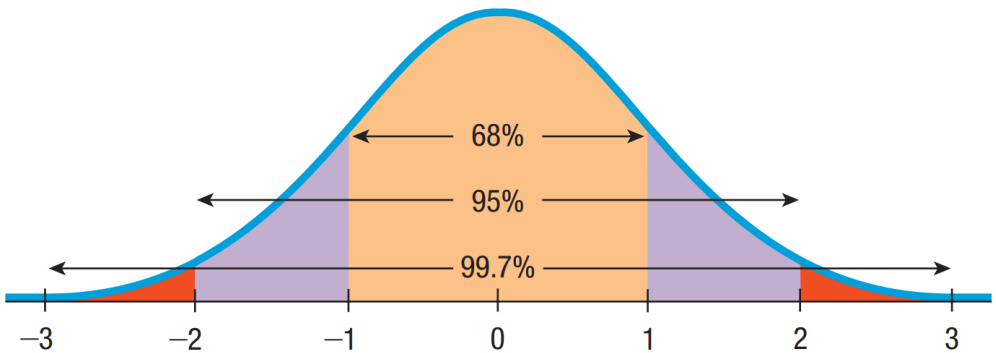


Stockholm University

Normalfördelningen - räkneexempel

Räkneexempel 1: Hur stor andel av hoppen är antingen minst två standardavvikelser större eller minst två standardavvikelser mindre än genomsnittet?

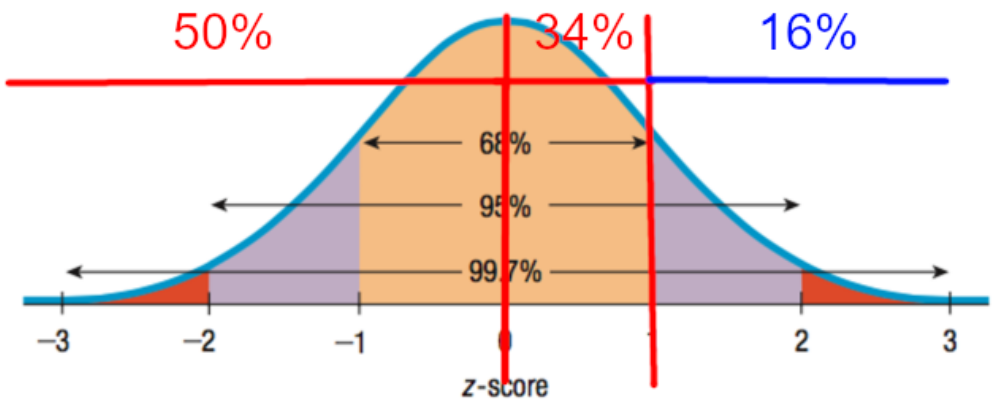
- Vi ser att om hopplängderna är normalfördelade så är 95% av hoppen i intervallet mellan -2 och 2 standardavvikelser från genomsnittet.
- Andelen hopp vars längd ligger utanför detta intervall är alltså $100\% - 95\% = 5\%$.



Normalfördelningen - räkneexempel

Räkneexempel 2: Hur stor andel av hoppen är minst en standardavvikelse högre än genomsnittet?

- Normalfördelningen är symmetrisk, så om vi delar den på mitten har vi 50% av alla hopp till vänster om mitten. De hopp som ligger på höger sida mellan 0 och 1 standardavvikelse utgör $68\%/2 = 34\%$ av alla hopp.
- Andelen hopp som är minst en standardavvikelse längre än genomsnittet är alltså $100\% - 50\% - 34\% = 16\%$.

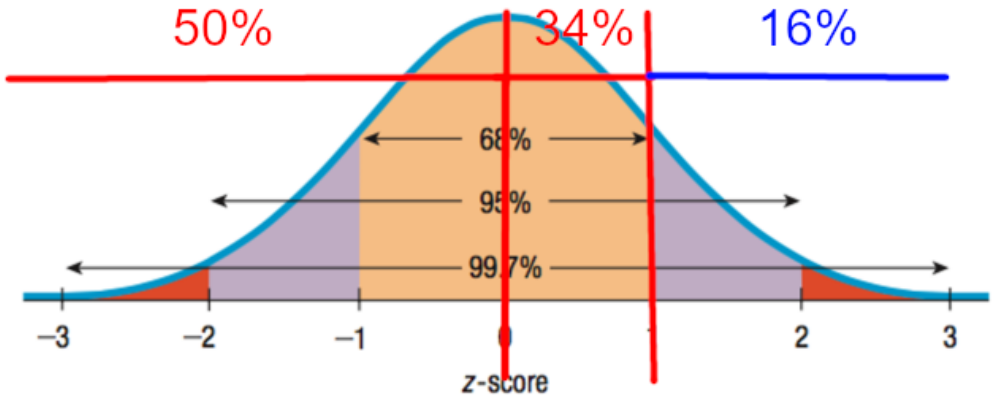


Stockholm University

Normalfördelningen - räkneexempel

Räkneexempel 3: Hur långt måste ett hopp minst vara **i meter** för att vara bland de 16% längsta hoppen?

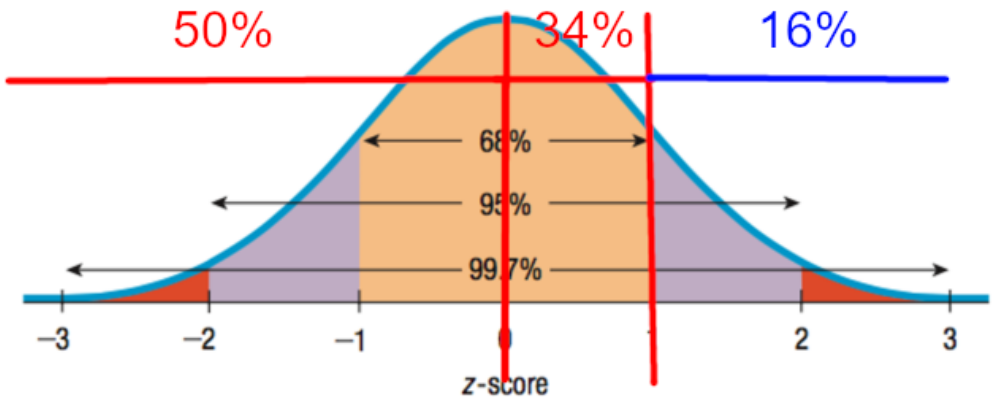
- Anta att genomsnittshoppet är $\bar{y} = 4.2$ meter, och standardavvikelsen är $s = 0.4$ meter.
- Det här är ett mer intressant exemepel än de tidigare. I vanliga fall tänker vi i termer av vanliga enheter (meter, kg, kronor, etc) och inte i termer av antalet standardavvikelser.



Normalfördelningen - räkneexempel

Räkneexempel 3: Hur långt måste ett hopp minst vara **i meter** för att vara bland de 16% längsta hoppen? $\bar{y} = 4.2$ meter, och $s = 0.4$ meter.

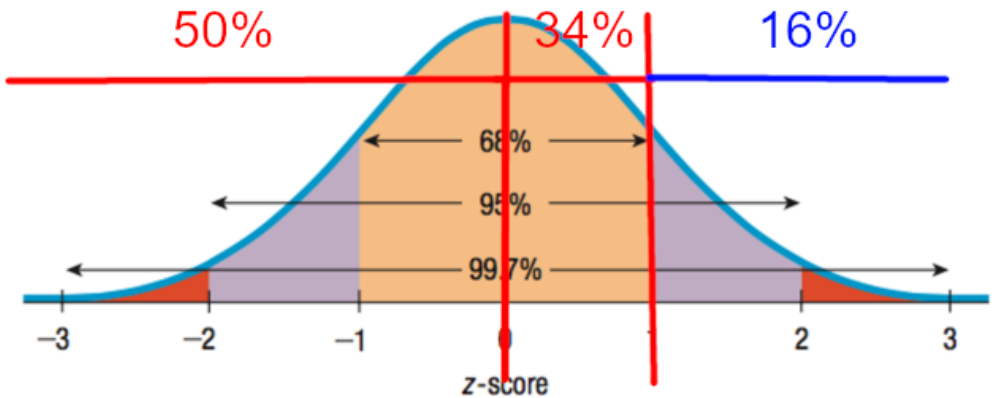
- För att vara bland de 16% längsta hoppen måste hoppet vara minst 1 standardavvikelse längre än genomsnittet.



Normalfördelningen - räkneexempel

Räkneexempel 3: Hur långt måste ett hopp minst vara **i meter** för att vara bland de 16% längsta hoppen? $\bar{y} = 4.2$ meter, och $s = 0.4$ meter.

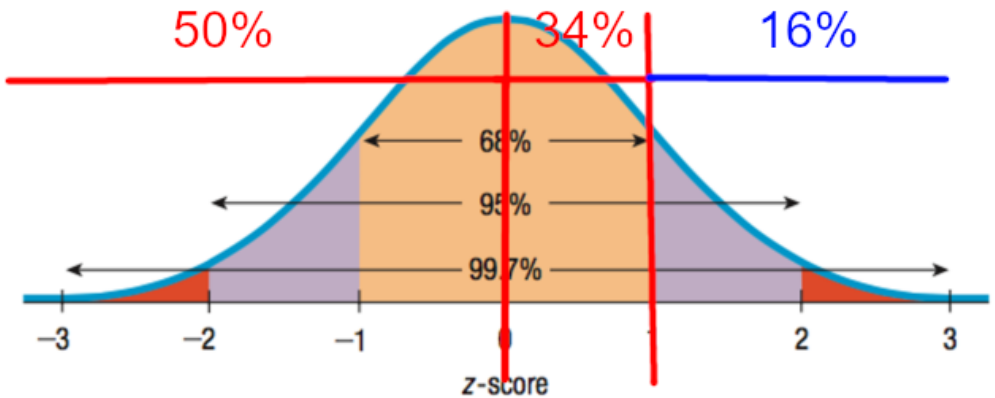
- För att vara bland de 16% längsta hoppen måste hoppet vara minst 1 standardavvikelse längre än genomsnittet.
- Det antalet meter som du måste hoppa är alltså $\bar{y} + 1 \cdot s$.



Normalfördelningen - räkneexempel

Räkneexempel 3: Hur långt måste ett hopp minst vara **i meter** för att vara bland de 16% längsta hoppen? $\bar{y} = 4.2$ meter, och $s = 0.4$ meter.

- För att vara bland de 16% längsta hoppen måste hoppet vara minst 1 standardavvikelse längre än genomsnittet.
- Det antalet meter som du måste hoppa är alltså $\bar{y} + 1 \cdot s$.
- $\bar{y} + 1 \cdot s = 4.2 + 1 \cdot 0.4 = 4.6$ meter.



Stockholm University

Normalfördelningen - percentiler

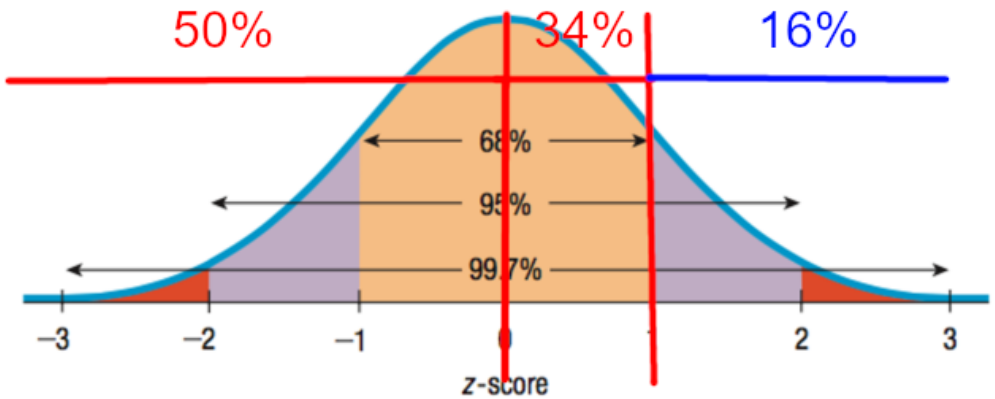
I föreläsning 2 nämndes **percentiler** lite kort. Om vi har en numerisk variabel så är en percentil ett värde som är större än en viss specificerad procentandel av observationerna.

Exempel

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 Q3.

Normalfördelningen - percentiler, exempel

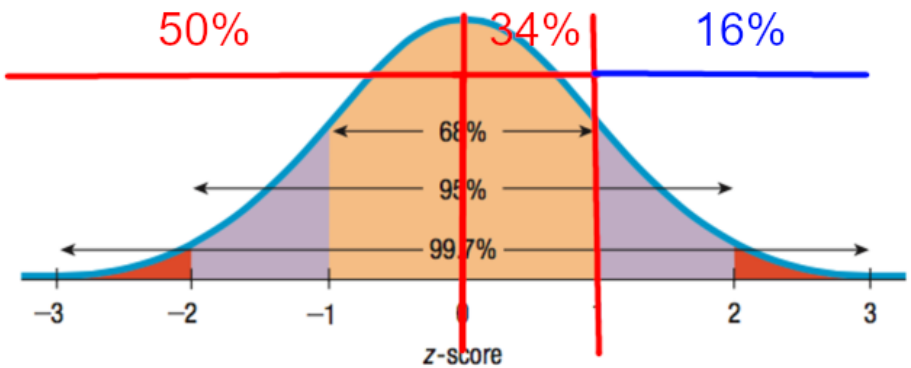
- Vi räknade ut att i en viss längshoppstävling måste ett hopp vara minst 4.6 meter för att ligga precis **en standardavvikele över genomsnittet**.
- I bilden nedan ser vi att ett hopp som är precis en standardavvikelse längre än genomsnittet så är hoppet längre än ungefär 84% av hoppen i tävlingen och kortare än ungefär 16% av hoppen. Ett hopp på 4.6 meter ligger alltså vid den **84:e percentilen**, förutsatt att hoppen är normalfördelade.



Stockholm University

Normalfördelningen - percentiler

- I verkligheten är behöver vi oftast mer information än vad bilden nedan ger.
- Anta att vi vill veta hur långt ett hopp i tävlingen minst måste vara för att tillhöra de 10 procent längsta hoppen, dvs ligga över den 90:e percentilen?
- Bilden visar att hoppet måste vara mer än 1 standardavvikelse över snittet (84:e percentilen), men det behöver inte vara så långt som 2 standardavvikelser över genomsnittet (97.5:e percentilen).
- För ett mer exakt svar behöver vi en **normalfördelningstabell**.

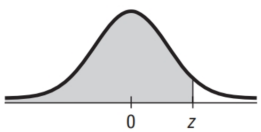


Stockholm University

Normalfördelningstabeller

- Normalfördelningstabellen, som vi här ser en liten del av, kan användas för att **översätta z-värden till proportioner, eller proportioner till z-värden**.
- Ett z-värde får du genom att kombinera talen i den vänstra marginalen med talen i den övre marginalen.
- För varje z-värde anger tabellen en proportion 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

Normalfördelningstabeller, att läsa tabellen

Låt oss titta på den markerade cellen. Den finns på en rad där det står 0.3 i vänstermarginalen och i en kolumn där det står 0.02 i den övre marginalen. Talen 0.3 och 0.02 sätts samman till 0.32. Talet i den övre marginalen anger alltså den andra decimalen i det z-värde som den markerade cellen avser.

					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

Normalfördelningstabeller, att läsa tabellen

Det gulmarkerade talet är 0.6255. Det betyder att vid z-värdet 0.32 har vi 62.55% av våra observationer till vänster om oss på normalfördelningskurvan. 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

Normalfördelningstabeller, tolkning

Om vi vill forsätta att prata om vår längdhoppstävling kan vi säga att om någon gör hopp som är 0.32 standardavvikelser längre än genomsnittet, då är det hoppet längre än 62.55% av alla hopp i tävlingen.

				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

Normalfördelningstabeller

- Ofta är vi inte intresserade av z-värdet som sådant. Vi använder det som ett mellansteg när vi översätter mellan originalenheten på vår ursprungliga variabel (meter, kg, kronor, etc) och en proportion i procent.
- Om du har ett värde i originalskalan och vill veta andelen observationer som är mindre eller större än detta värde:

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

- Om du vill veta vad värdet i originalskalan behöver vara för att en bestämd andel av observationerna ska vara mindre/större:

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

- Låt oss göra en uträkning av varje slag!

Normalfördelningstabeller, räkneexempel 1

- I vår längdhoppstävling gör vi ett hopp på $y = 4.5$ meter. Genomsnittshoppet är $\bar{y} = 4.2$ meter, och standardavvikelsen är $s = 0.4$ meter.
- Vi vill nu veta andelen hopp som är kortare respektive längre än vårt hopp, så vi gör våra uträkningar i den här ordningen:

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

- z-värdet räknar vi som vanligt ut med formeln $z = (y - \bar{y})/s = (4.5 - 4.2)/0.4 = 0.75$. Vårt hopp var alltså 0.75 standardavvikelser längre än genomsnittshoppet.

Normalfördelningstabeller, räkneexempel 1

- Vi har räknat ut att vårt hopp var 0.75 standardavvikelser längre än genomsnittshoppet. Z-värdet är alltså 0.75.
- För att se andelen hopp som har ett **lägre** z-värde än 0.75 använder vi normalfördelningstabellen. På raden med 0.7 i sidomarginalen och i kolumnen med 0.05 i toppmarginalen hittar vi 0.7734 (Kontrollera gärna själv att du får samma resultat).
- **Slutsats:** Ungefär 77.34 procent av alla hopp i tävlingen är kortare än vårt hopp, vilket betyder att 22.66 procent av hoppen är längre.
- Vi kan, om vi vill, säga att ett hopp på 4.5 meter befinner sig ungefär vid den 77:e percentilen i förhållande till övriga hopp i tävlingen.

Normalfördelningstabeller, räkneexempel 2

- Vårt senaste hopp var inte bland de 10 procent längsta. Hur långt måste vi hoppa om vi har det målet? Det gäller fortfarande att $\bar{y} = 4.2$ meter, $s = 0.4$ meter.
- Den här gången är andelen känd. Vi måste hoppa så långt att minst 90 procent av hoppen i tävlingen är kortare än vårt hopp.
- Detta är samma sak som att vårt hopp måste ligga över den 90:e percentilen. Vi gör våra uträkningar i den här ordningen:

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

Normalfördelningstabeller, räkneexempel 2

- Den här gången letar vi efter andelen 0.9 i normalfördelningstabellen. Den exakta andelen 0.9 finns inte i tabellen, så vi tar det närmaste värdet som är 0.8997.
- Andelen 0.8997 finns på en rad där det står 1.2 i sidomarginalen och i en kolumn där det står 0.08 i toppmarginalen. Vi sätter ihop detta till 1.28. Nu vet vi att vårt hopp måste vara minst 1.28 standardavvikelser längre än genomsnittet för att vara bland de 10 procent längsta.
- Då återstår att omvandla z-värdet till meter. Vi använder formeln $y = \bar{y} + zs = 4.2 + 1.28 \cdot 0.4 = 4.712$.
- **Slutsats:** Vi måste hoppa längre än 4.712 meter för att vårt hopp ska vara bland de 10 procent längsta.

Normalfördelningstabeller

Med funktionerna *qnorm* och *pnorm* i R kan du hitta samma information som i normalfördelningstabellen.

- *pnorm* använder du när du har ett z-värde och vill veta hur stor andel av observationerna i en normalfördelning som har ett lägre z-värde. I ett av våra längdhoppsexempel ville vi veta hur stor andel av hoppen som hade ett lägre z-värde än 0.75. I tabellen såg vi att det var 77.34 procent.
- Vi

```
pnorm(0.75)
```

```
[1] 0.7733726
```

Normalfördelningstabeller

Med funktionerna *qnorm* och *pnorm* i R kan du hitta samma information som i normalfördelningstabellen.

qnorm använder du för att se vilket z-värde som är större än en bestämd procentandel av observationerna. I tabellen såg vi att det z-värde som var större än 90 procent av observationerna, och mindre än övriga 10 procent, var 1.28.

```
qnorm(0.9)
```

```
[1] 1.281552
```

Notera att vi i R inte skriver andelarna i procent.

Varför just normalfördelningen?

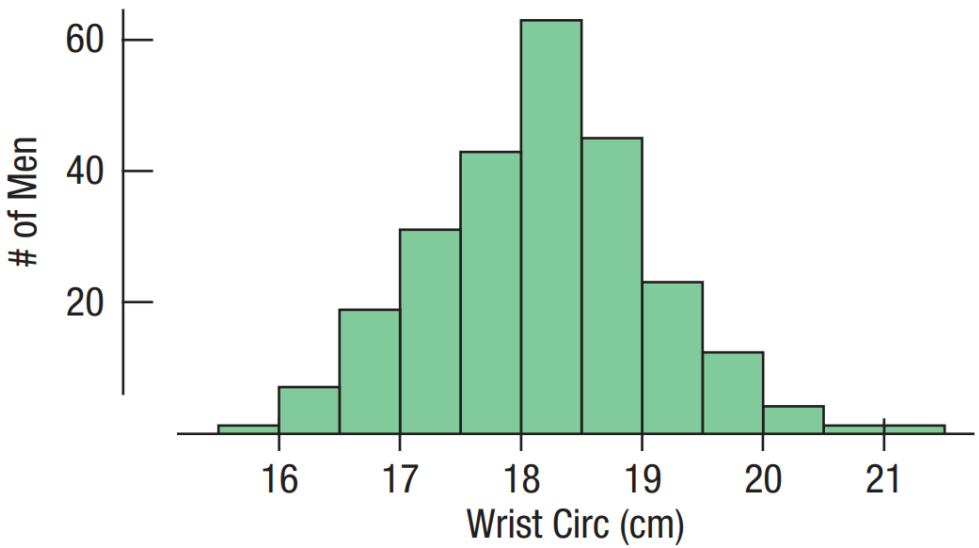
- En numerisk variabel vara fördelad på många olika sätt. Varför är vi så intresserade av just variabler som är normalfördelade?
- Den franske matematikern Abraham de Moivre visade på 1700-talet att många fördelningar i verkligheten faktiskt ligger nära normalfördelningen (som dock inte hade fått sitt namn på den tiden).
- Dessutom är det så att **medelvärden** fördelar sig enligt en normalfördelning. Det här 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 de beräkningar vi har gjort i våra räkneexempel.
- Vi kan inte utan argument utgå från att en numerisk variabel är normalfördelad.
- När vi använder normalfördelningen för våra beräkningar måste vi alltså först visa att vår variabel verkligen är normalfördelad, eller åtminstone att den följer en fördelning som liknar en normalfördelning.

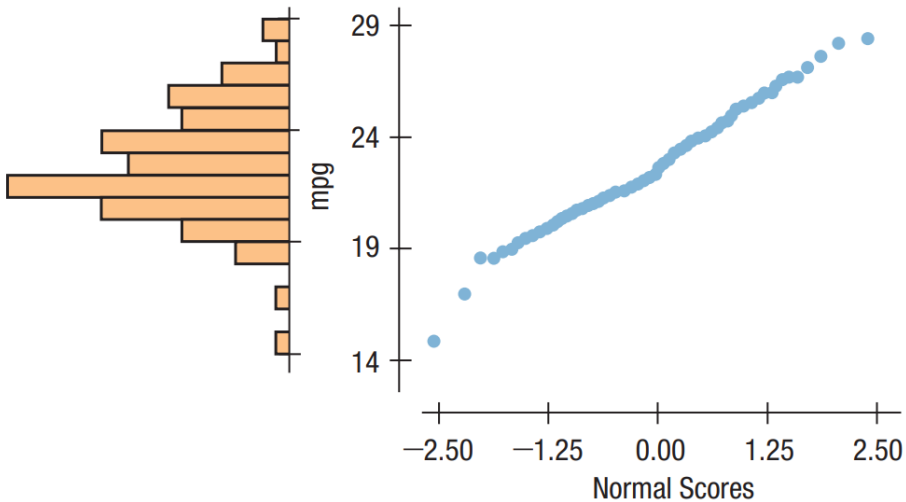
Undersök om en variabel är normalfördelad

- Vi kan säga att en fördelning är **nästan normalfördelad (nearly normal)** om den är symmetrisk, bara har en topp (unimodal) och inte har uppenbara outliers.
- Fördelningen på bilden får sägas leva upp till villkoren för att vara nästan normalfördelad. 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 ligger punkterna i plotten längs en rät linje.
- **Exempel:** Bensinförbrukning i mpg (miles per gallon) för en Nissan Maxima insamlat 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.



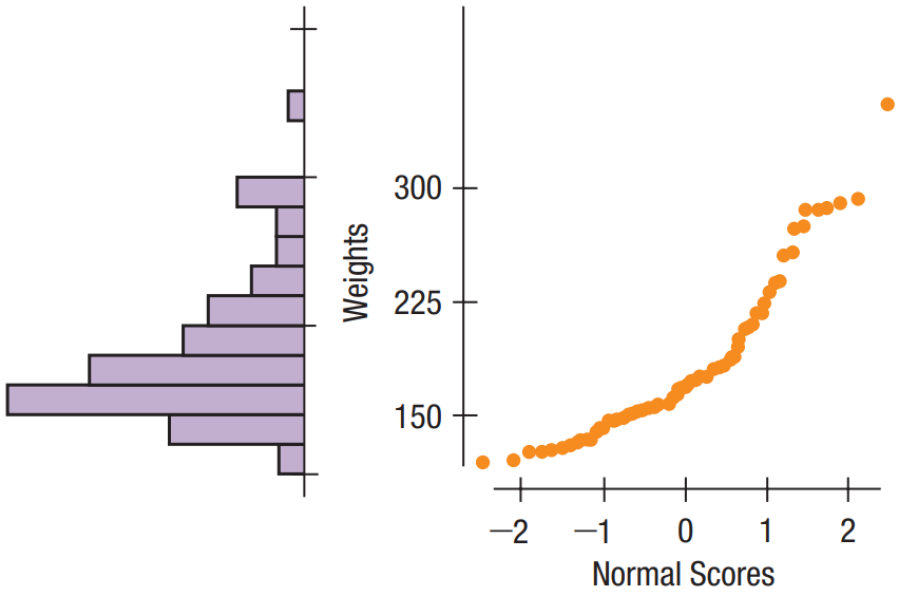
Undersök om en variabel är normalfördelad

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.

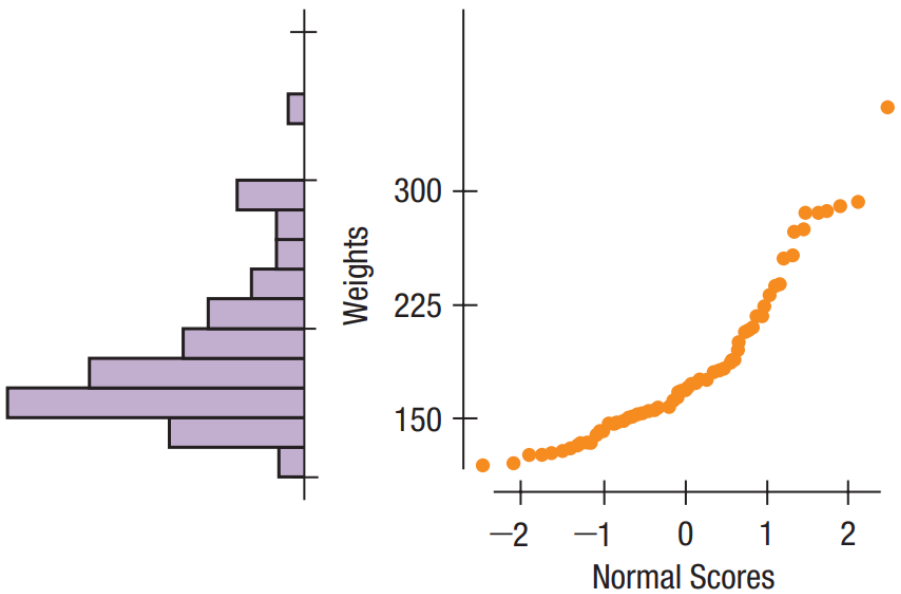
Undersök om en variabel är normalfördelad

- Här är en normalfördelningsplot och ett histogram ur De Veaux et al. (2021).
- Linjen är inte rak. Dessutom ser vi i histogrammet att fördelningen inte är symmetrisk, utan skev åt höger. Det vore orimligt att betrakta variabeln som normalfördelad.



Undersök om en variabel är normalfördelad

- Det kan vara möjligt att göra en variabel som denna normalfördelad genom en transformation.
- Vi såg tidigare att transformationer kan göras för att underlätta jämförelser mellan variabler med låddiagram. Att göra en variabel normalfördelad är också en vanlig anledning att göra en transformation.



Stockholm University

Undersök om en variabel är normalfördelad

I R kan vi använda bland annat funktion *hist* för att göra ett histogram. Vi kan sedan använda och *qqnorm* och *qqline* tillsammans för att göra en normalfördelningsplot, med en linje som visar hur punkterna bör ligga.

```
data(mtcars)
par(mfrow=c(1, 2))
hist(mtcars$mpg, breaks=15)
qqnorm(mtcars$mpg)
qqline(mtcars$mpg)
```