Föreläsning 4

karl.sigfrid@stat.su.se

Vad har vi gjort hittills

- Hittills har vi tittat på
  - fördelningar av kategoriska variabler (stapeldiagram)
  - fördelningar av numeriska variabler (histogram)
  - samband mellan kategoriska variabler (korstabeller)
- Nu är det dags att undersöka hur samband kan se ut mellan en kategorisk och en numerisk variabel.

Repetition

- När vi intresserar oss för **kategoriska variabler** vill vi oftast veta hur stora olika grupper av observationer är.
- När vi intresserar oss för **numeriska variabler** kan vi vilja undersöka **centralmått** som medelvärde och median.
- För numeriska variabler kan vi också vilja undersöka **spridningsmått** som , standardavvikelse och kvartilavstånd (Inter Quartile Range) (se föreläsning 2).

Numeriska variabler betingade på kategoriska

- När vi vill undersöka sambandet mellan en numerisk variabel och en kategorisk variabel kan vi studera fördelningen av den numeriska variabeln **betingat** på den kategoriska variabeln.

Numeriska variabler betingade på kategoriska

- När vi vill undersöka sambandet mellan en numerisk variabel och en kategorisk variabel kan vi studera fördelningen av den numeriska variabeln **betingat** på den kategoriska variabeln.
- Med **två kategoriska variabler** kan vi ställa frågor som: Är andelen BMW-förare som kör för fort större än andelen Toyota-förare som kör för fort?

Numeriska variabler betingade på kategoriska

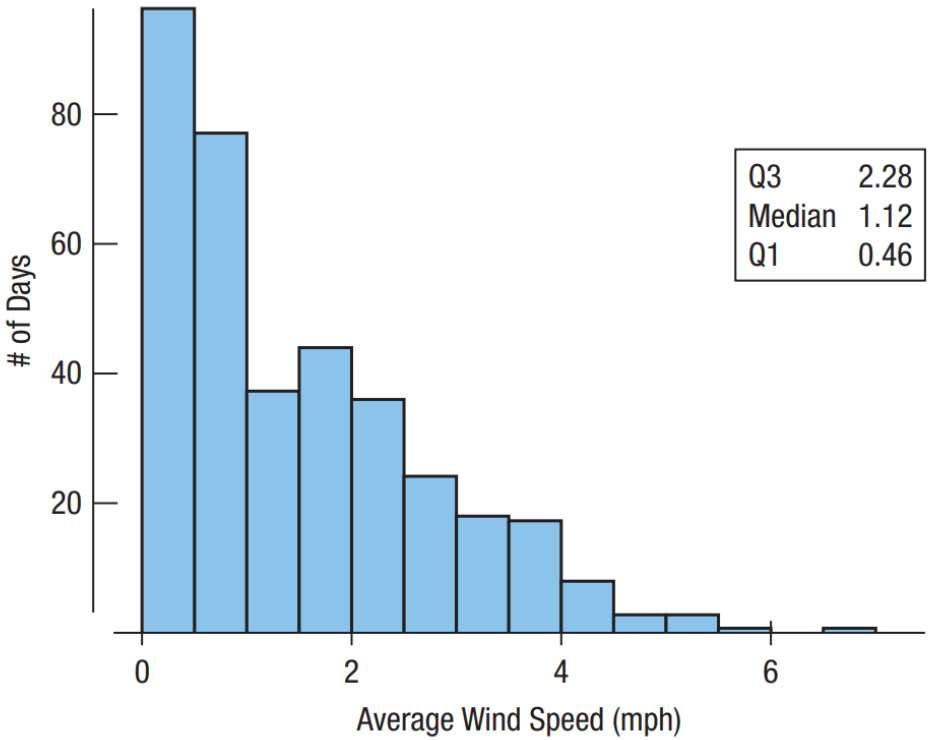
- När vi vill undersöka sambandet mellan en numerisk variabel och en kategorisk variabel kan vi studera fördelningen av den numeriska variabeln **betingat** på den kategoriska variabeln.
- Med **två kategoriska variabler** kan vi ställa frågor som: Är andelen BMW-förare som kör för fort större än andelen Toyota-förare som kör för fort?
- Med **en numerisk och en kategorisk variabel** kan vi ställa frågor som: Hur snabbt kör BMW-förare i genomsnitt på en viss vägsträcka? Som jämförelse, hur snabbt kör Toyota-förare i genomsnitt på samma vägsträcka?

Numeriska variabler betingade på kategoriska

- När vi vill undersöka sambandet mellan en numerisk variabel och en kategorisk variabel kan vi studera fördelningen av den numeriska variabeln **betingat** på den kategoriska variabeln.
- Med **två kategoriska variabler** kan vi ställa frågor som: Är andelen BMW-förare som kör för fort större än andelen Toyota-förare som kör för fort?
- Med **en numerisk och en kategorisk variabel** kan vi ställa frågor som: Hur snabbt kör BMW-förare i genomsnitt på en viss vägsträcka? Som jämförelse, hur snabbt kör Toyota-förare i genomsnitt på samma vägsträcka?
- Vi undersöker alltså hastigheten (numerisk variabel) betingat på bilmärket (kategorisk variabel).

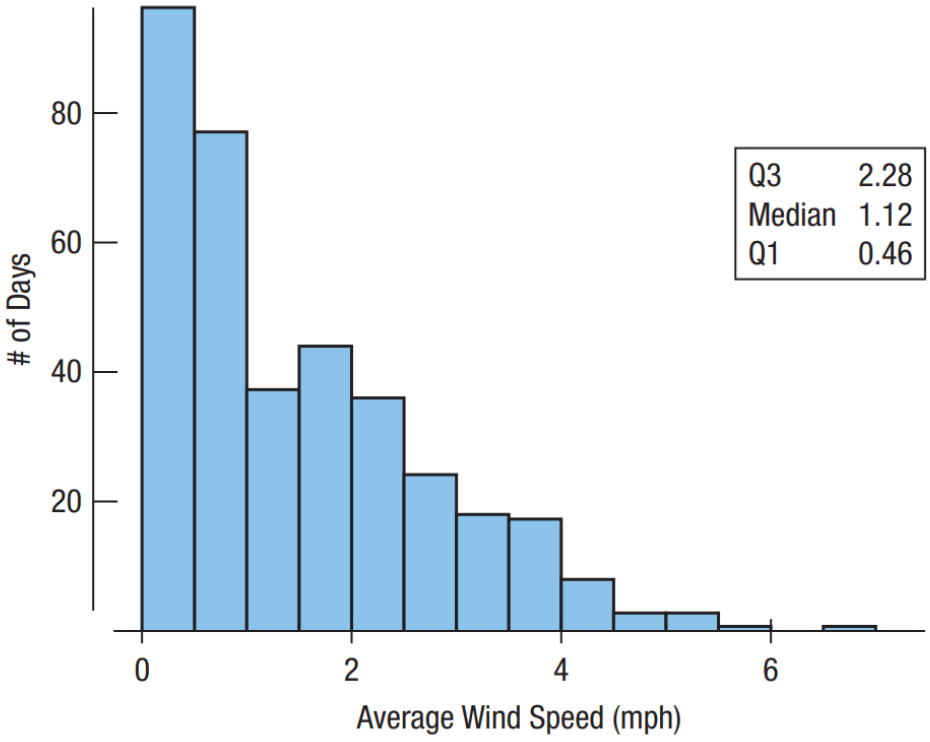
Varför betinga numeriska variabler på kategoriska

Figur 4 i De Veaux et al. (2021) visar den dagliga medelvindhastigheter under 2011 i western Massachusetts. För hälften observationerna är vindhastigheten mindre än 1.12 mph. Fördelningen är skev till höger.



Varför betinga numeriska variabler på kategoriska

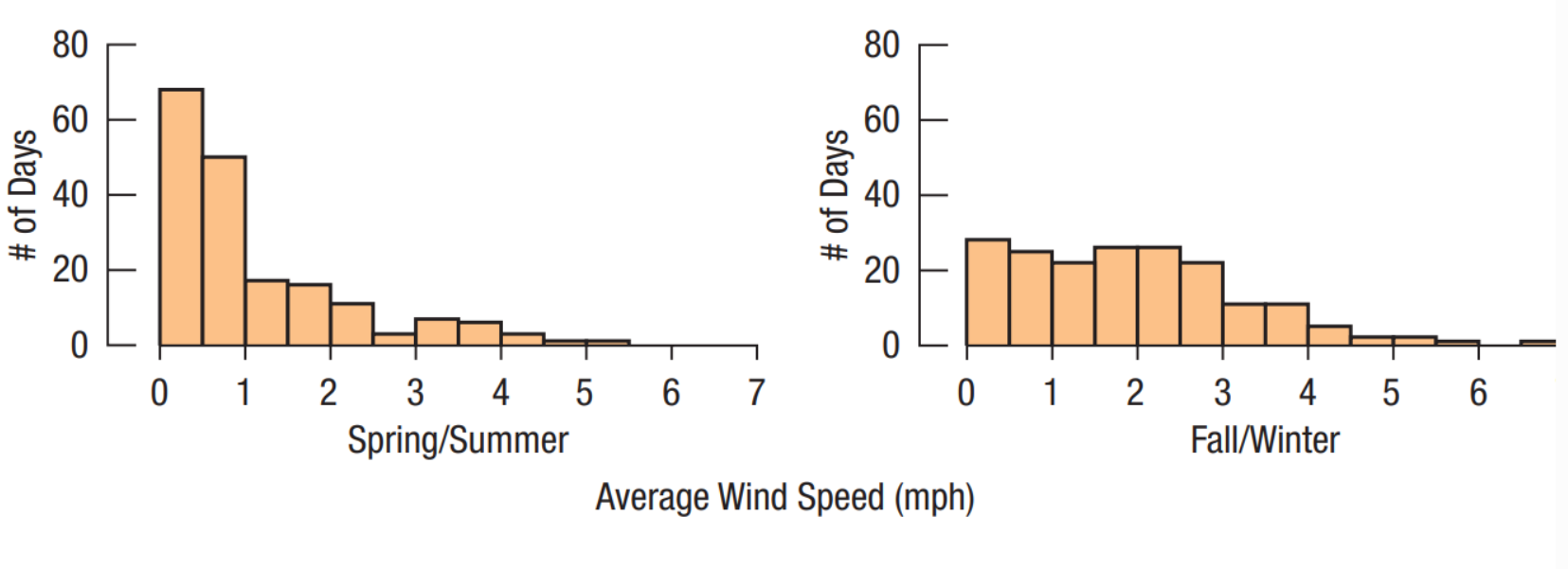
Fråga: Är den här fördelningen av vindhastigheter representativ för alla delar av året? Det kan vi vinte veta utifrån detta histogram.



Stockholm University

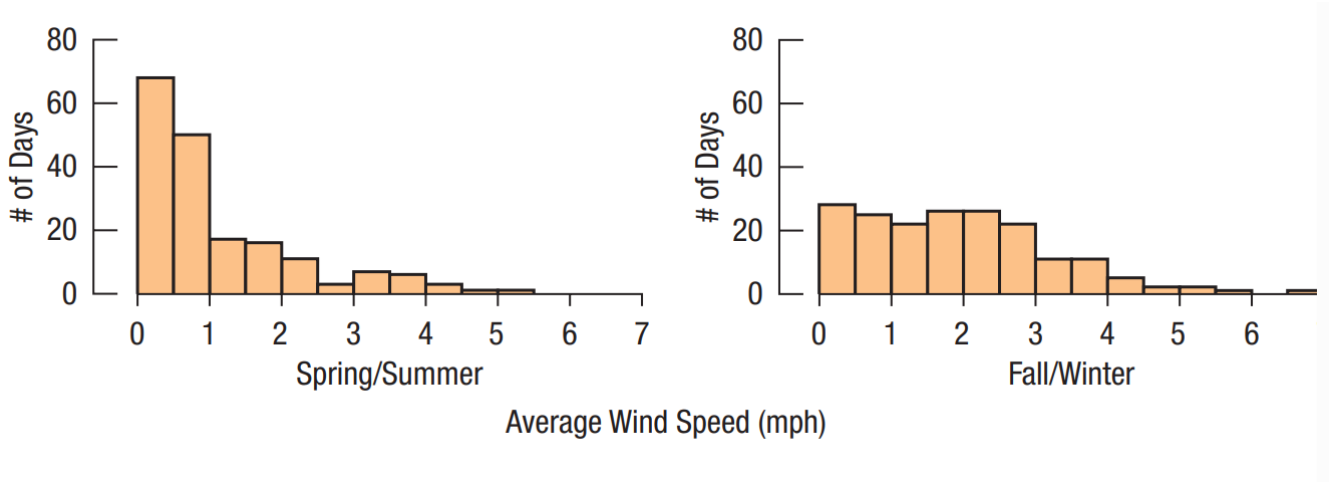
Varför betinga numeriska variabler på kategoriska

Figur 4.2 i De Veaux et al. (2021) visar medelvindhastigheten separat för två säsonger: vår/sommar och höst/vinter.



Varför betinga numeriska variabler på kategoriska

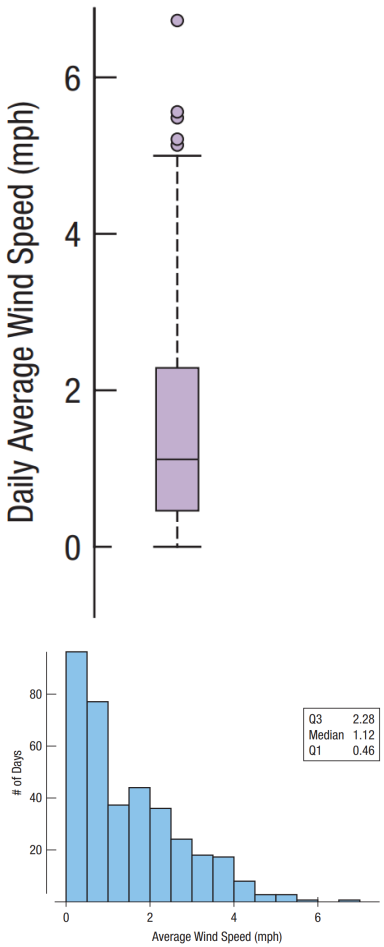
- Om vi jämför del fördelningen för hela året kan vi se att det är mindre vind under vår/sommar, och det är fler dagar med mycket vind under höst/vinter.
- Är vi intresserade av vindstyrkan under en viss tid på året ger de betingade fördelningarna en bättre bild än marginalfördelningen. (Från föreläsning 3: Marginalfördelningen är den fördelning av en variabel som inte tar hänsyn till andra variabler).



Varför betinga numeriska variabler på kategoriska?

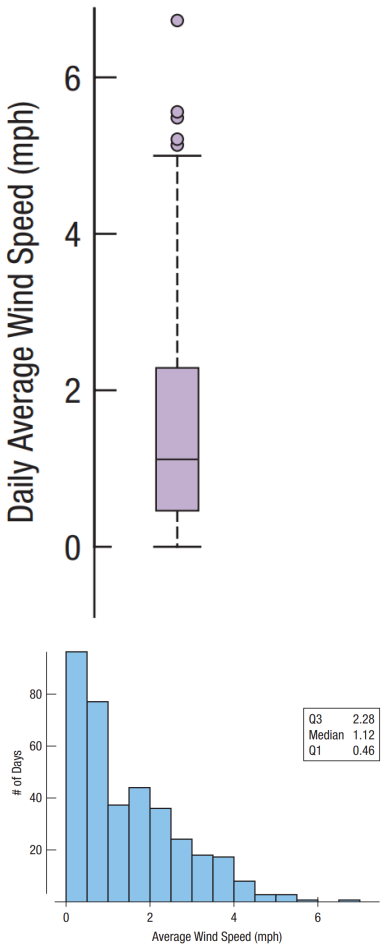
- Vi har nu sett att vi får en bättre bild av vindstyrkan om vi betingar variabeln på säsong.
- Vi kan få en ännu bättre bild av vindstyrkan om vi bryter ned observationerna i 12 månadsgrupper istället för bara två säsonger.
- Men hur ska vi illustrera de tolv månaderna? Tolv separata histogram blir svåröverskådligt.
- Som ett alternativ till histogram kan vi använda **låddiagram (box plot)**.

Låddiagram (boxplot)



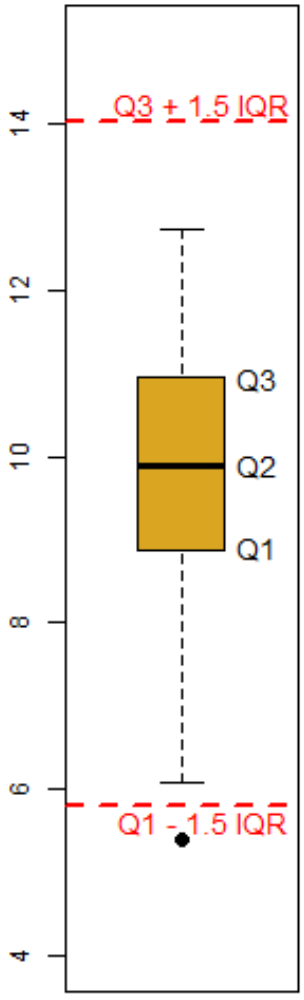
- Den övre bilden till vänster föreställer ett låddiagram. Låddiagrammet visar fördelningen av dagar med olika vindstyrka i Western Massaschussets.
- Låddiagrammet Ilustrerar samma data som som det blå histogrammet, men på ett mer sammanfattande sätt.

Låddiagram (boxplot)



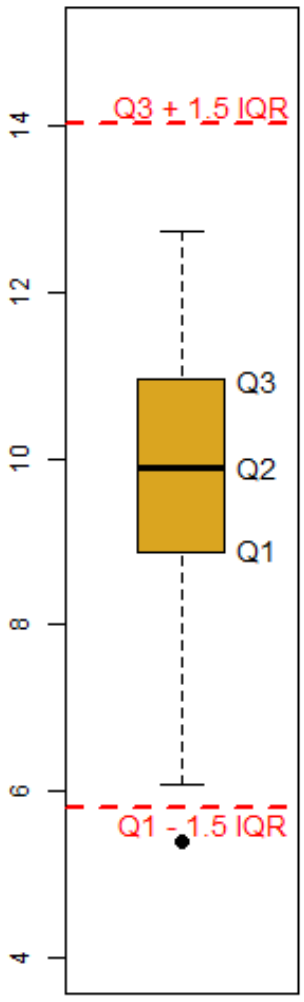
- Medan histogrammet ger en mer komplett bild av fördelningen visar låddiagrammet ett antal nyckelmått:
  - Medianen
  - Q1
  - Q3
  - Kvartilavståndet
  - Värdet av den största och den minsta observationen

Låddiagram (boxplot) - hur det är uppbyggt



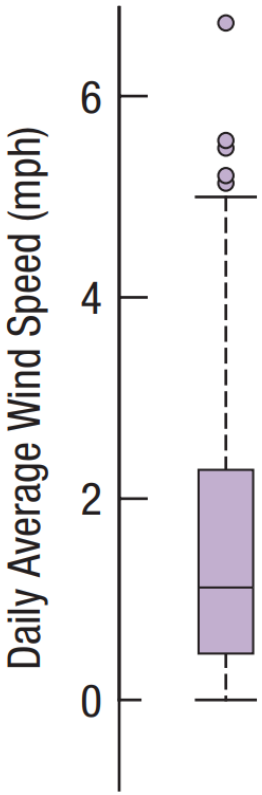
- Figuren består av en **låda (box)**, **morrhår (whiskers)** och **en punkt**.
- Undre kanten av lådan mäter **Q1**.
- Övre kanten av lådan mäter **Q3**.
- Linjen som går genom lådan mäter **medianen (Q2)**.
- Lådans höjd mäter **kvartilavståndet (IQR)**.

Låddiagram (boxplot) - hur det är uppbyggt



- **De röda stödlinjerna** är inte en del av diagrammet. De är placerade vid $Q1 - 1.5 \cdot IQR$ respektive $Q3 + 1.5 \cdot IQR$.
- **Morrhåren** sträcker sig till det minsta respektive största värde som ligger innanför de röda stödlinjerna.
- Eventuella observationer som ligger utanför stödlinjerna räknas som **outliers** och ritas ut som punkter. I detta låddiagram har vi en sådan punkt under den nedre röda linjen. Vi har inga outliers i den över delen av diagrammet.
- Diagrammet kan också vara liggande.

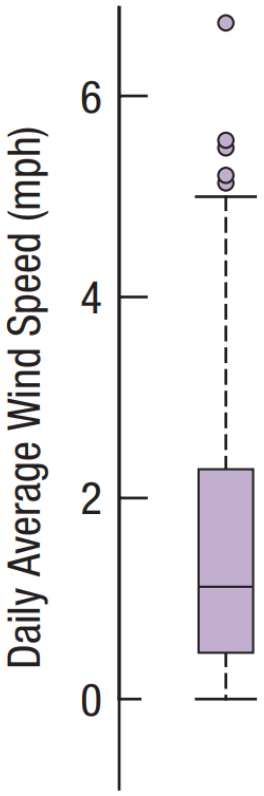
Låddiagram - tolkning



Vi använder det vi vet om låddiagram för att utläsa information om vindstyrkan i Western Massaschussets:

- Värdet för Q1 är lite över 0 mph.
- värdet för Q2 omkring 1 mph.
- värdet för Q3 lite över 2 mph.
- Det lägsta värdet är omkring 0 mph.
- Det högsta vädet ligger en bit över 6 mph.

Låddiagram - tolkning

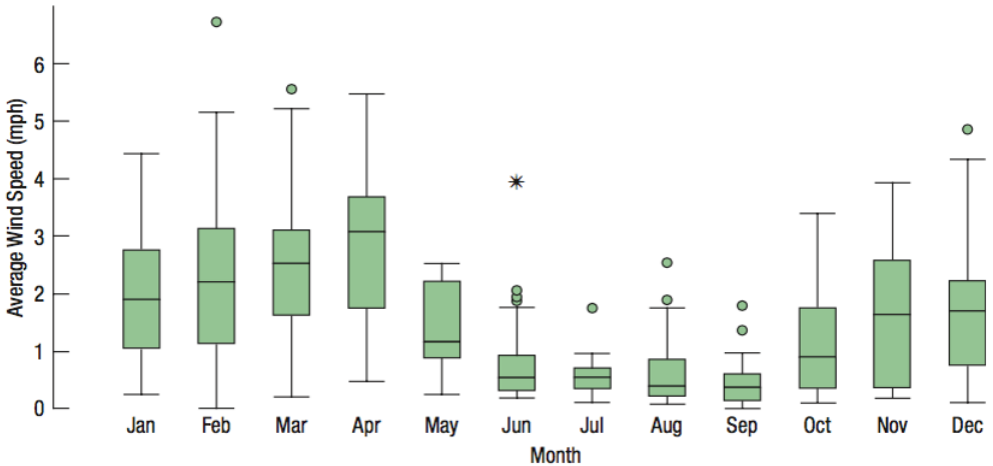


Vi använder det vi vet om låddiagram för att utläsa information om vindstyrkan i Western Massaschussets:

- Medianen ligger närmare Q1 än Q3 och det övre morrhåret är längre än det undre morrhåret. Det talar för att fördelningen är skev åt höger.
- Det finns ett antal höga outliers, men inga låga outliers.
- Obs! Att observationer identfieras som outliers behöver inte betyda att de ska tas bort! Vi bör dock vara medvetna om dem.

Jämför fördelningar med låddiagram

En fördel med låddiagram jämfört med histogram är att flera fördelningar enkelt kan jämföras. Figur 4.3 i De Veaux et al. (2021) visar hur vindstyrkorna i Massaschussets betingade på månad. Vi ser ett tydligt mönster.



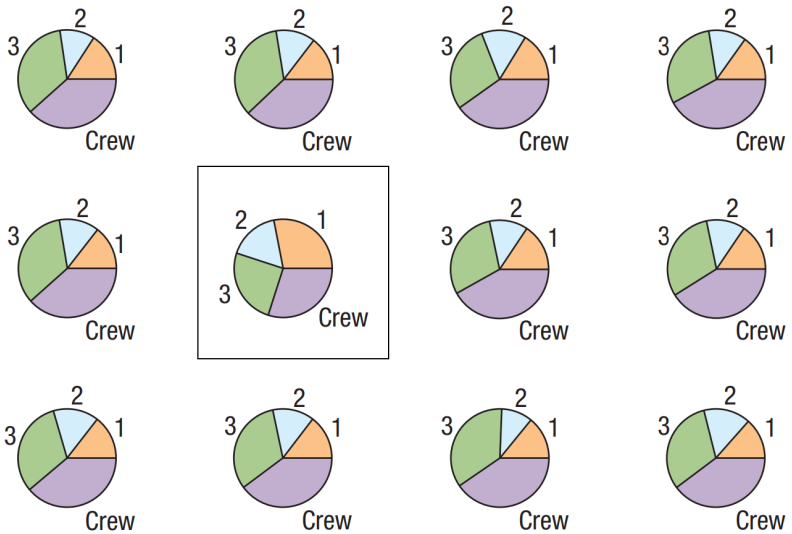
- Det är mer vindstilla och mindre variation under sommaren.
- Notera stjärnan över juni. Denna extrema outlier representerar en tornado.

Numeriska betingade fördelningar - samband och slump

- På föreläsning 3 tog vi upp frågan om **skillnader mellan grupper** i ett datamaterial, och hur vi kan bedöma om skillnaderna beror på slumpen eller på att det finns ett mer generellt samband.
- I exemplet på föreläsning 3 tittade vi på hur Titanics livbåtar fördelades mellan passagerare som reste i olika biljettklasser.
- Vi ställde upp **hypotesen** att möjligheten att få plats i en livbåt var slumpmässig och oberoende av biljettklass.

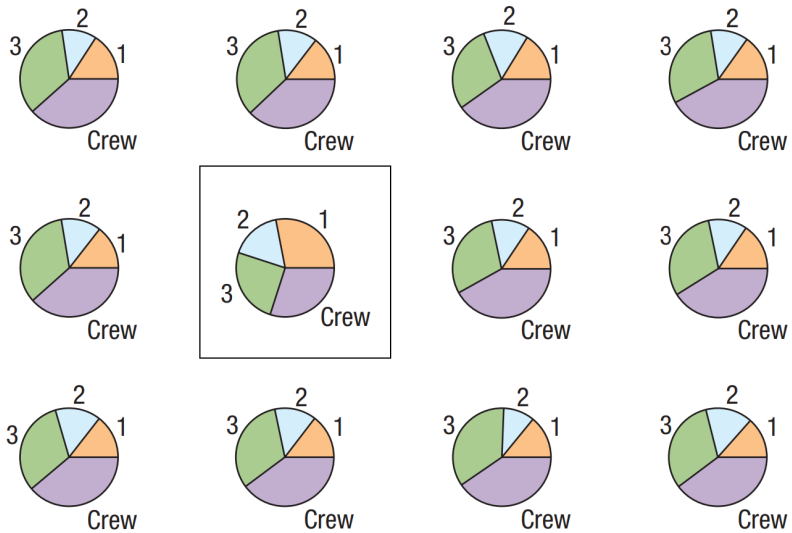
Numeriska betingade fördelningar - samband och slump

- Vi upprepade ett experiment där vi lät platserna i livbåtarna fördela sig slumpvis mellan alla passagerare, **som om vår hypotes var sann**. För varje upprepning gjorde vi ett pajdiagram.

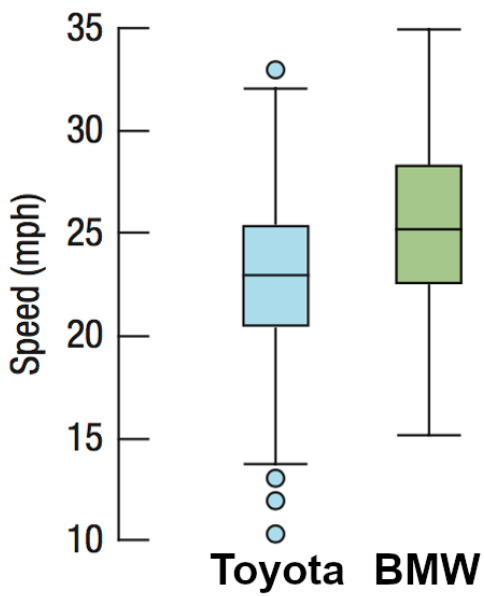


Numeriska betingade fördelningar - samband och slump

- Vi undersökte om den verkliga fördelningen av platserna i livbåtarna verkade troligt **givet att hypotesen är sann**. Nu ska vi göra samma sak, med den här gången frågar vi oss om **skillnaden mellan två numeriska fördelningar** beror på slumpen eller inte.



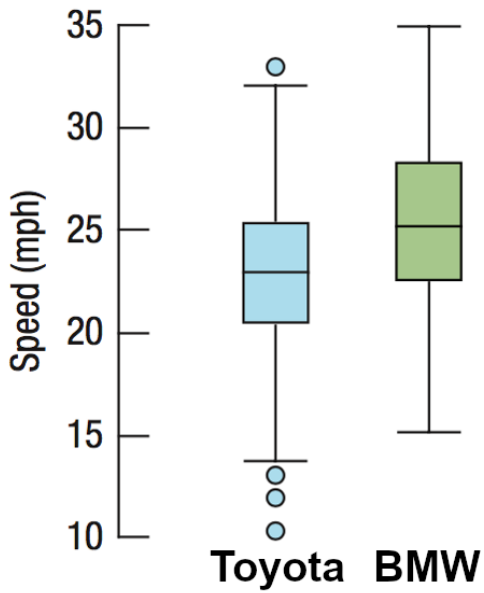
Numeriska betingade fördelningar - samband och slump



Vi mäter hastigheten på bilar som kör längs en gata. De två låddiagrammen visar hastighetsfördelningen för dem som kör BMW-bilar och för dem som kör Toyota-bilar.

- Medelhastigheten hos de bilar som kör BMW är 2.53 mph högre än medelhastigheten hos de som kör Toyota.
- Betyder det att BMW-förare generellt kör snabbare än Toyota-förare, eller beror skillnaden i medelhastighet på slumpen?

Numeriska betingade fördelningar - samband och slump



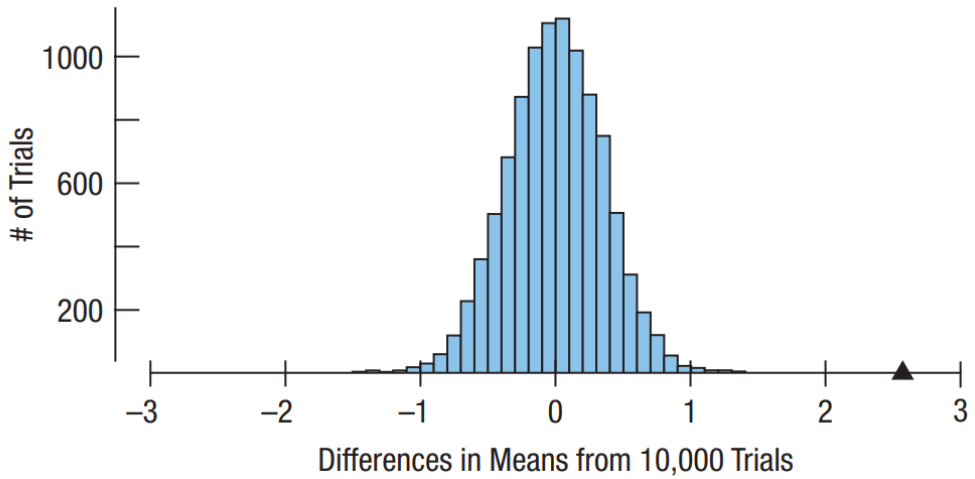
- Vi vill veta om skillnaden i medehastighet beror på på att de som kör det ena bilmärket generellt kör snabbare, eller om de olika medehastigheterna är slumpen.
- Vi kan ställa upp **hypotesen** att hastigheten är oberoende av bilmärke.
- Om hypotesen är sann så berodde det enbart på slumpen att BMW-förarna körde snabbare i just de fall som vi mätte.

Numeriska betingade fördelningar - samband och slump

- **Vi gör ett tankeexperiment:** Anta att vi inte känner till vilket bilmärke som är kopplat till vilken hastighet i vår data.
- För varje hastighet slumpar vi fram fram om bilmärket är BMW eller Toyota. I vårt tankeexperiment blir nu hastigheten oberoende av bilmärket.
- Efter att varje hastighet i vårt dataset slumpvis har parats ihop med en bilmodell räknar vi ut medelhastigheten för vardera bilmodell.
- Vi noterar skillnaden i medelhastighet för de två grupperna. Vi räknar skillnaden som medelhastigheten för BMW-bilarna minus medelhastigheten för Toyota-bilarna.

Numeriska betingade fördelningar - samband och slump

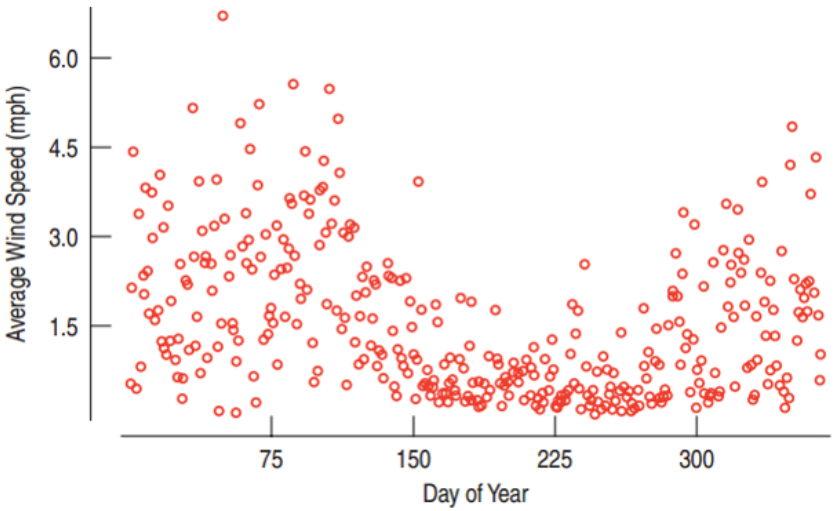
- Nu upprepar vi detta slumpexperiment **10,000 gånger**. Figur 4.5 i De Veaux et al. (2021) visar utfallet av 10,000 sådana experiment.
- Vilken slutsats om vår hypotes kan vi dra? Skillnaden i medelhastighet som uppmättes i studien, 2.53 mph, är markerad med en triangel.



Stockholm University

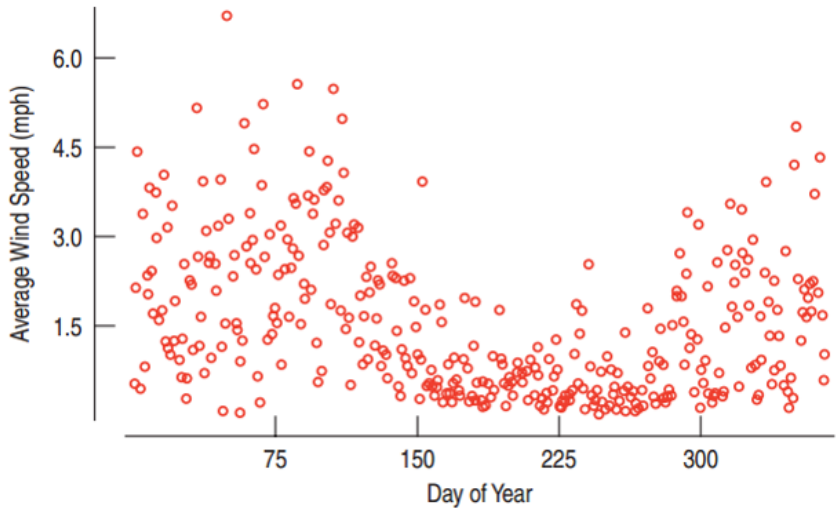
Spridningsdiagram (scatter plot)

- Vi har hittills tittat på diagram som på olika sätt sammanfattar numeriska värden, så som histogram och låddiagram.
- Ibland vill vi ha en bild som visar varje observation. Det kan vi åstadkomma med ett **spridningsdiagram (scatter plot)**.
- Figur 4.6 i De Veaux et al. (2021) visar medelvindstyrkan för varje dag 2011.



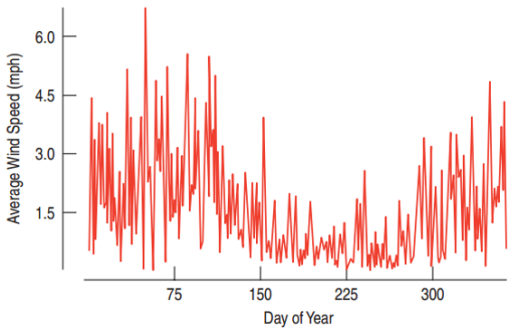
Tidsserier

- På **y-axeln** ser vi medelvindstyrkan i Western Massaschussets.
- På **x-axeln** ser vi hur många dagar in på året vi är. Observationerna är alltså ordnade i tidsordning från vänster till höger. Därmed illustrerar diagrammet en **tidsserie**.



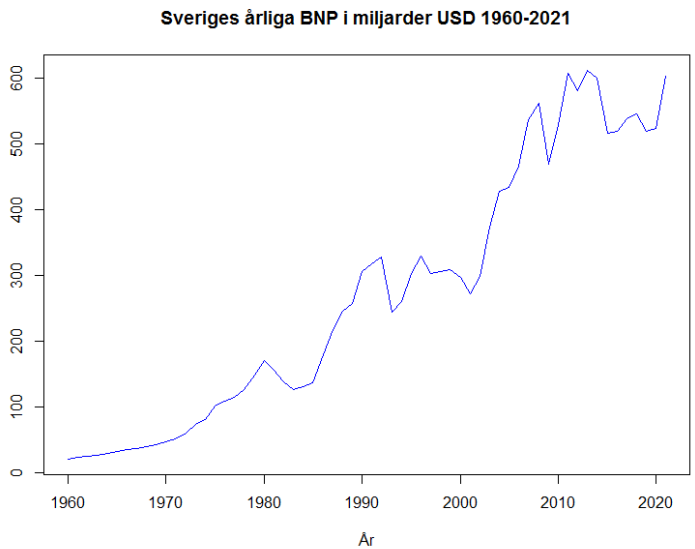
Tidsserier

- Ofta binder en tidsserieplot ihop punkterna med en linje som gör det lättare att se mönster.
- Intressant att titta efter är
  - Trender
  - Säsongsvariation
- Trender och säsongsvariationer kan vara viktiga om du vill göra framtidsprognoser.



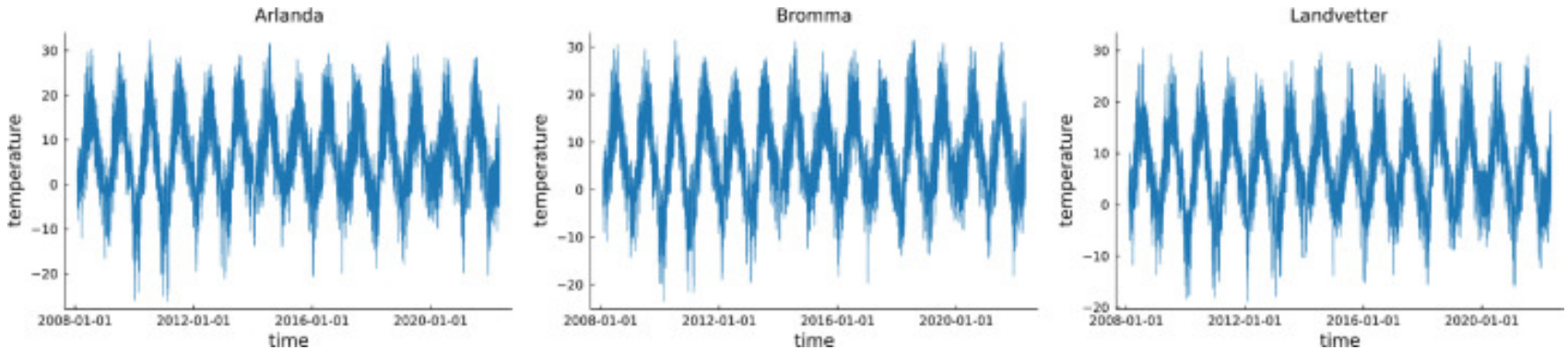
Tidsserier - trender

- **En trend** är en kontinuerlig förändring som sker över tid.
- Den här grafen visar Sveriges bruttonationalprodukt från 1960 till 2020.
- Även om BNP sjunker vissa år är trenden positiv. Om vi drog en rak linje från 1960 till 2020 skulle linjen peka kraftigt uppåt.



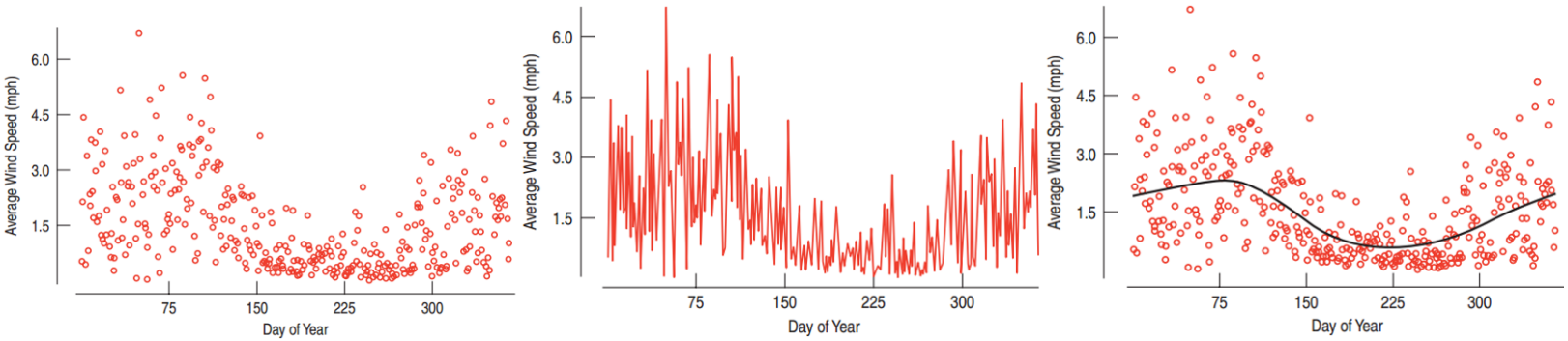
Tidsserier - säsongsvariation

- Exempel på tidsserier som har **säsongsvariation** är temperaturer.
- Bilden visar temperaturer insamlade mellan 1 februari 2008 och 1 maj 2022 vid tre svenska flygplatser. (Bild från Villani et al. (2022))



Tidsserier - utjämning (smoothing)

- Spridningsdiagrammet utan linjer till vänster, och med linjer i mitten är båda kaotiska.
- Den svarta linjen genom spridningsdiagrammet till höger är en utjämnad kurva.
- En av de enklaste metoderna för utjämning är att använda ett glidande medelvärde (moving average).



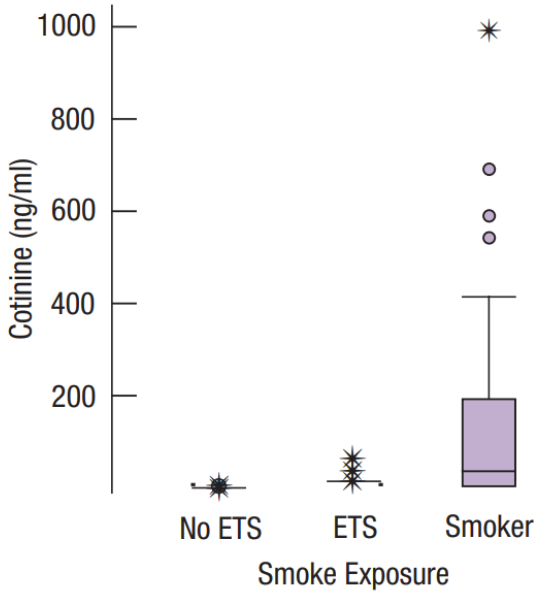
Transformationer

För att förstå varför man kan vilja **transformera** en variabel ska vi till på en studie som undersökte om exponering för rökning påverkade nivån av kotinin i blodet (nedbrytningsprodukt av nikotin).

- Deltagarna delade in i tre grupper:
  - Rökare (Smoker)
  - Passiva rökare (ETS)
  - De som inte utsattes för någon rök (No ETS).
- ETS står för Exposed to smoke.
- Vi har alltså en kategirisk variabel (grupptillhörighet) och en numerisk variabel (mängd kotonin i blodet).

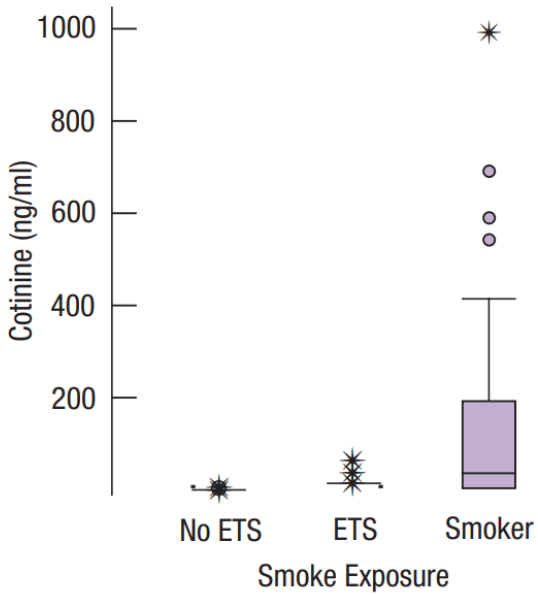
Transformationer

- Här är ett låddiagram för de tre grupperna.
- Ser vi några problem med diagrammet?



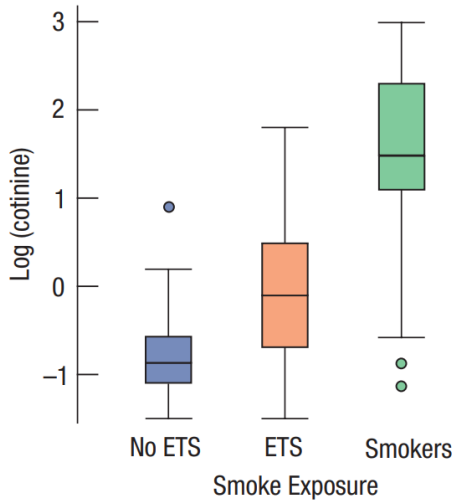
Transformationer

- När värden är ojämnt fördelade kan det svara svårt att läsa ett diagram. Här är en stor del av alla värden ihopklämda i botten av diagrammet.
- Det är omöjligt att se skillnaden mellan passiva rökare (ETS) och de som inte exponerats för rök (No ETS).



Transformationer

Om vi transformerar våra värden på y-axeln från cotinine till log(cotinine) får vi det här låddiagrammet:



Nu ser vi tydligt att de som exponerats för rök har högre halter kotinin i blodet än de som inte exponerats. Kom ihåg att det inte är mängden kotinin i blodet (i nanogram/ml) som vi mäter på y-axeln, utan det logaritmerade värdet.

Logaritmer

Följande samband är bra att känna till för att kunna översätta mellan logaritmer och vår ursprungliga skala:

$$y = e^x \iff \log(y) = x$$

Vi kan också skriva

$$y = e^{log(y)}$$

Uttrycket e är en konstant med ett värde som är ungefär 2.7.

Exempel

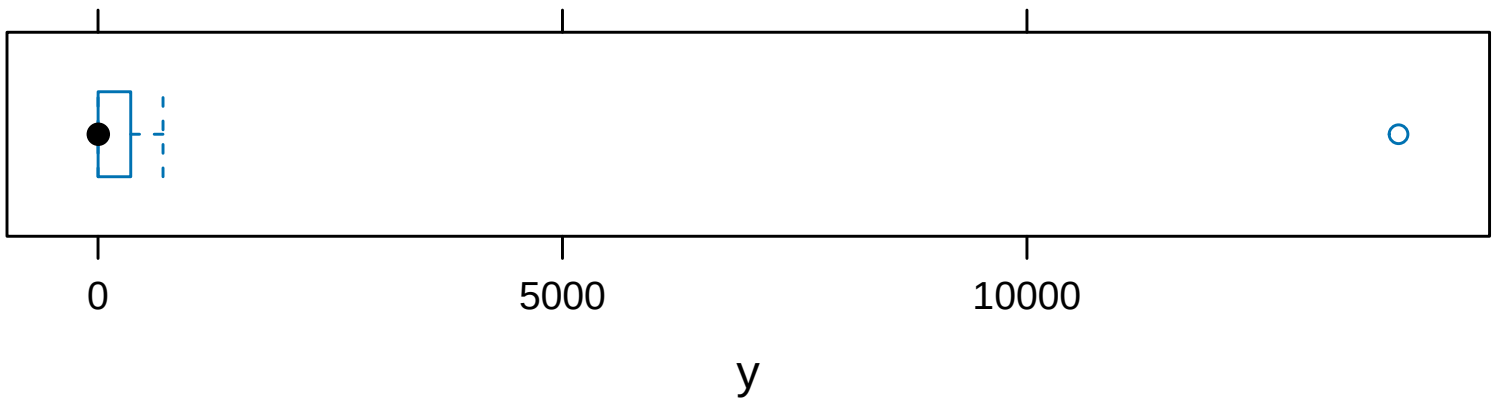
Anta att vi vet att $\log(a) = 1.2$, och vi vill ha värdet av a. Vi vet då att $a = e^{\log(a)}$, vilket innebär att

$$a = e^{\log(a)} = e^{1.2} = 3.32$$

Transformationer i R

Vi skapar en variabel som vi kallar y, med värden som skiljer sig kraftigt i storleksordning. Vi försöker illustrera y med ett låddiagram.

```
y <- c(0.2, 0.3, 2, 3,  5, 700, 14000) #Skapa en vektor
bwplot(y) # Gör en boxplot av y med mosaic-paketet
```



Resultatet blir ett låddiagram som är svårt att läsa. Nästan alla observationer är ihoptryckta i diagrammets vänstra del.

Transformationer i R

Vi skapar variabeln logy, som är logaritmen av y, och skriver ut den nya variabelns värden avrundade till 3 decimaler.

```
logy <- log(y)
print(round(logy, 3)) # Skriv ut den logaritmerade vektorn,
```

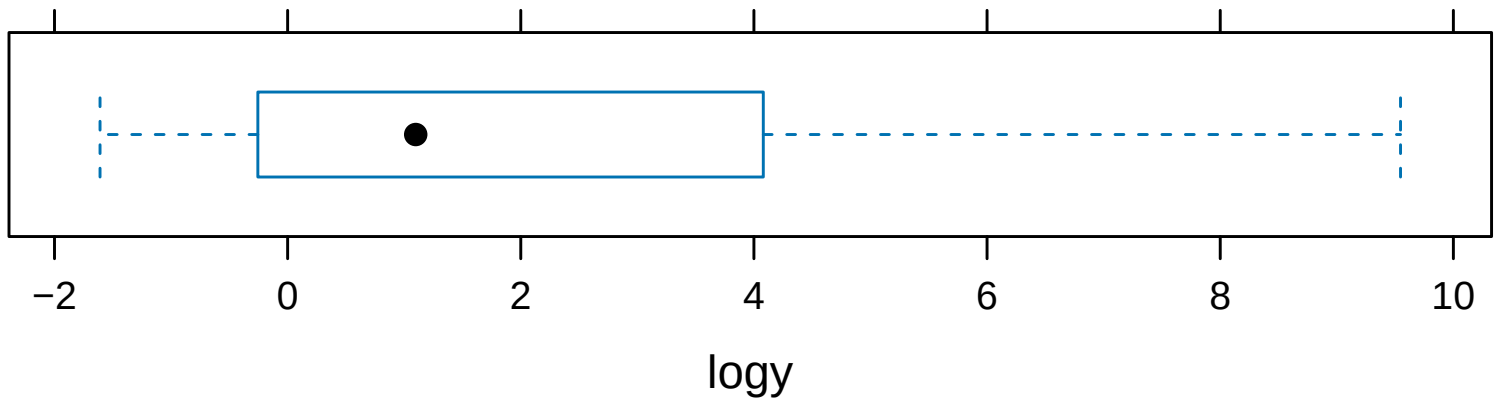
```
[1] -1.609 -1.204  0.693  1.099  1.609  6.551  9.547
```

Trots att våra ursprungliga värden varierar stort, från 0.2 till 14 000, håller sig våra logaritmerade värden inom ett intervall från ungefär -1.6 till 9.5.

Transformationer i R

Vi sammanfattar våra värden i logy med ett låddiagram.

```
bwplot(logy) # Gör en boxplot av y med mosaic-paketet
```



Resultatet blir ett låddiagram som är mer lättläst.

Transformationer i R

- Vi kan transformera tillbaka våra värden till orginalskalan med formeln $y = e^{\log(y)}$, där $e \approx 2.718$.
- Notera att e^x i R skrivs *exp(x)*.

```
y_backtransformed <- exp(logy) # Transformera tillbaka
y_backtransformed # Skriv ut y
```

```
[1]     0.2     0.3     2.0     3.0     5.0   700.0 14000.0
```

Vi ser att de värden som transformerats tillbaka är våra ursprungliga värden.

Transformationer

- Logaritmering är bara en av många transformationer som vi kan ha nytta av.
- När vi kommer till avsnittet om regression kommer transformering att spela en större roll.

Om du tycker att transformationer verkar jobbigt, oroa dig inte! När vi gör transformationer på den här kursen kommer det inte att handla om matematik, utan om att pröva sig fram.