Statistik och dataanalys I

F4: Jämföra fördelningar, tidsserier, och transformationer.

Valentin Zulj

Statistiska institutionen

Stockholm
University

Vad har vi gjort hittills

- Hittills har vi pratat om
  - Fördelningar för kategoriska variabler (t.ex. stapeldiagram)
  - Fördelningar för numeriska variabler (t.ex. histogram)
  - Hur vi illustrerar/presenterar samband mellan kategoriska variabler (t.ex. korstabeller)

Vad vi ska göra idag

- Samband mellan en kategorisk och en numerisk variabel
  - Betinga numeriska variabler på kategoriska
  - Lådagram
  - Samband och slump
  - Spridningsdiagram
- Tidsserier (väldigt kort, mer i SDA II)
  - Säsongsvariation
  - Utjämning
- Transformationer

Betinga numeriska variabler på kategoriska

Repetition

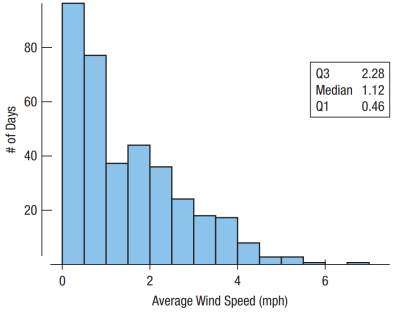
- När vi intresserar oss för **kategoriska variabler** vill vi oftast veta hur stora olika grupper av observationer är
  - T.ex. hur många som har en viss diet som får cancer, hur många som har läst en viss utbildning som får jobb kort efter examen, etc
- När vi intresserar oss för **numeriska variabler** kan vi vilja undersöka
  - **centralmått** som medelvärde och median (någon typ av genomsnitt, t.ex. *hur många månader tar det att få jobb efter examen, i genomsnitt?*)
  - **spridningsmått** som, standardavvikelse och kvartilavstånd (se F2, t.ex. *hur stor är spridningen i tid till sysselsättning efter examen?*)
- När det gäller kategoriska variabler är det ofta "bara" att räkna, men det är inte lika entydigt hur vi svarar på de senare frågorna
- Vad betyder *i genomsnitt*, och hur definierar vi *spridning*?

Numeriska variabler betingade på kategoriska

- För att 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 motsvarande för Toyota-förare?
- 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) betingad (givet, uppdelad) på bilmärket (kategorisk variabel)

Varför betinga på detta sätt?

- Varför vill vi ens dela upp analyser på detta vis? Ibland kan
  - Jämförelse av grupper vara det som driver vår analys (t.ex. tid till sysselsättning för nyexade vars föräldrar har/inte har eftergymnasial examen)
  - Gemensamma analyser ibland ge felaktiga bilder av läget i respektive grupp
- Figur 4 i De Veaux et al. (2021) visar medelvindhastigheter i västra Massachusetts (dagliga observationer under 2011)
- Hälften av observationerna är mindre än 1.12 mph, och fördelningen är skev till höger



Varför betinga på detta sätt?

- **Fråga:** Är den här fördelningen representativ för alla delar av året?
- Detta kan vi inte avgöra utifrån histogrammet ovan, utan vi behöver då dela upp analysen efter (betinga på) årstid

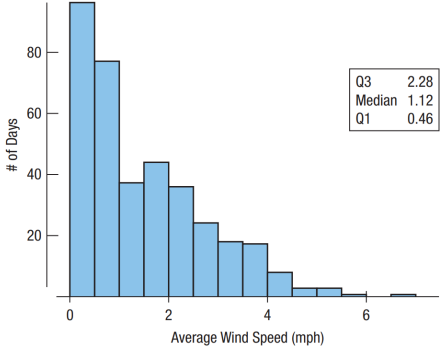
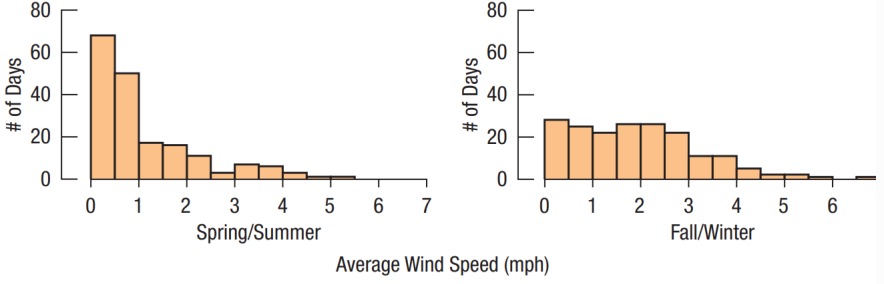
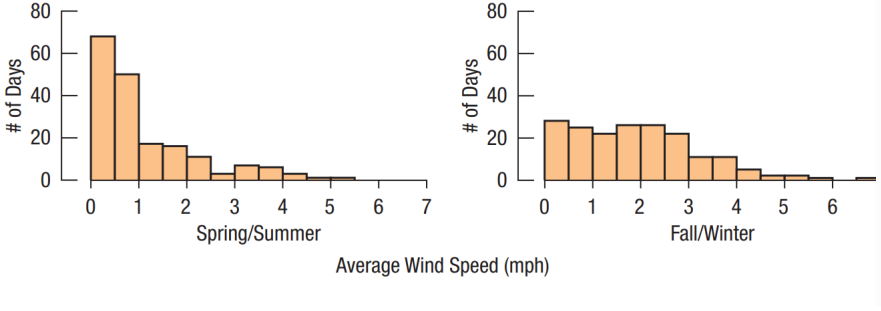


Illustration av betingad fördelning



- Figur 4.2 i De Veaux et al. (2021) visar medelvindhastigheten separat för två halvår: vår/sommar och höst/vinter
- Jämfört med fördelningen för hela året kan vi se att
  - Det blåser mindre under vår/sommar
  - Det är fler dagar med mycket vind under höst/vinter

Illustration av betingad fördelning

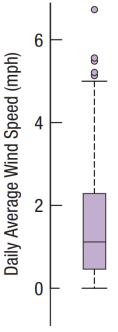


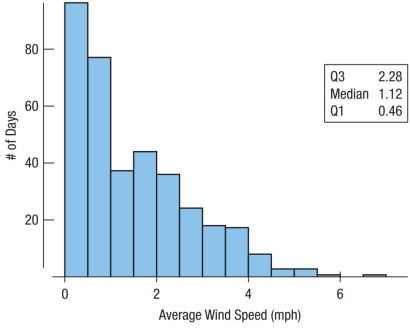
- Är vi intresserade av vindstyrkan under en viss tid på året ger de betingade fördelningarna en bättre bild än marginalfördelningen
- Från föreläsning 3: Marginalfördelningen är den fördelning av en variabel som inte tar hänsyn till andra variabler (t.ex årstid)

Varför betinga numeriska variabler på kategoriska?

- Vi får en bättre bild av vindstyrkan om vi betingar variabeln på årstid
- Vi kan få en ännu bättre bild av vindstyrkan om vi bryter ned observationerna i 12 månadsgrupper istället för bara två halvår
- Men hur ska vi illustrera de tolv månaderna? Tolv separata histogram blir svåröverskådligt
- Som ett alternativ till histogram kan vi använda **låddiagram/lådagram** (*en: boxplot*)

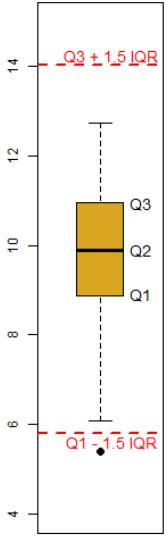
Låddiagram





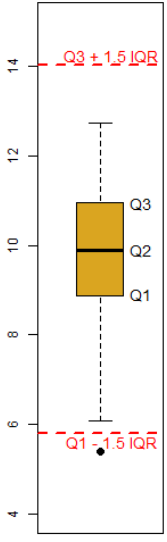
- Figurerna ovan visar båda fördelningen av dagliga vindstyrkor i västra Massachussets
- Figuren till vänster visar ett **lådagram**, och den till höger ett vanligt histogram
- Histogrammet ger en mer en mer övergripande bild av fördelningen
- Lådagrammet innehåller med detaljerad information om median, kvartiler, kvartilavstånd, och minsta/största observation

Detaljerna i ett låddiagram

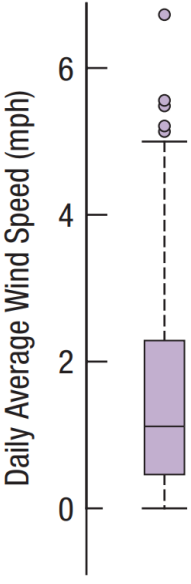


- Figuren består av en **låda (box)**, **morrhår (whiskers)** och **en punkt**
- Lådans nedre gräns visar $\mathbf{Q_1}$
- Lådans övre gräns visar $\mathbf{Q_3}$
- Linjen som går genom lådan visar **medianen** (dvs $\mathbf{Q_2}$)
- Lådans höjd (avstånd från Q_1 till Q_3) ger **kvartilavståndet (IQR)**
- **Morrhåren** sträcker sig till det minsta respektive största värdet som ligger innanför de röda stödlinjerna

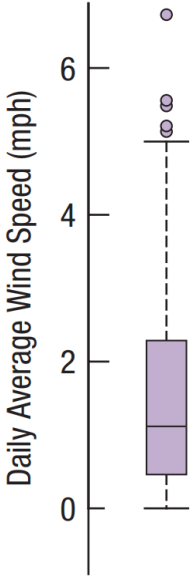
Låddiagram - hur det är uppbyggt



- **De röda stödlinjerna** är inte en del av diagrammet. De är placerade vid $Q1 - 1.5 \cdot IQR$ respektive $Q3 + 1.5 \cdot IQR$
- Observationer som ligger utanför stödlinjerna räknas som **outliers** och ritas ut som punkter
- I detta fall har vi en sådan punkt under den nedre gränsen
- Vi har inga outliers i den över delen av diagrammet
- Diagrammet kan också vara liggande



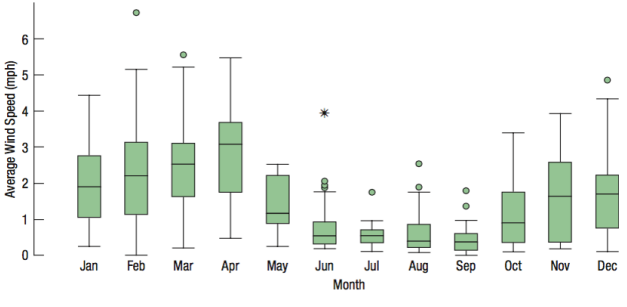
- Vi kan få nedan information ur lådagrammet med vindhastigheter
- Q_1 är lite större än 0 mph
- Q_2 är omkring 1 mph
- Q_3 är lite större än 2 mph
- Det lägsta värdet är omkring 0 mph
- Det högsta vädet är en bit över 6 mph
- **Fundering:** Kan vi avgöra om fördelningen är skev eller symmetrisk?



- Vi ser att medianen ligger närmare Q1 än Q3, och det övre morrhåret är längre än det undre
- Detta talar för att fördelningen är skev åt höger (observationerna är mer utspridda åt höger)
- Det finns ett antal höga outliers, men inga låga outliers
- **Obs:** Observationer som identfieras som outliers bör ej tas bort utan vidare! Borttagning måste motiveras noggrant

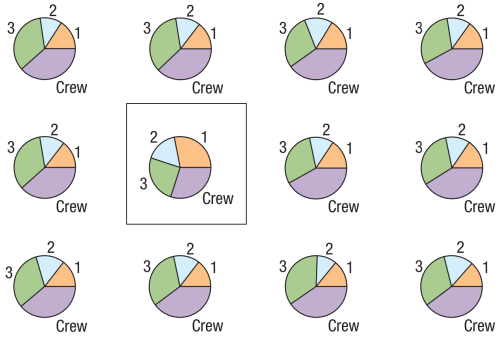
Jämföra fördelningar med låddiagram

- Lådagram är små och kompakta, jämfört med histogram, och kan användas för att beskriva/jämföra flera olika fördelningar
- Det är enkelt att skapa en figur med t.ex. 12 lågadram bredvid varandra, om vi vill jämföra vindhastigheter för varje enskild månad – som Figur 4.3 i De Veaux et al. (2021)



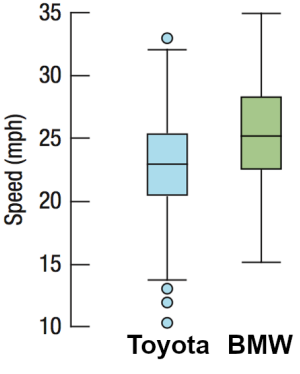
- **Mönster:** det blåser mindre under sommaren, med mindre variation
- **Stjärnan** över juni är en **extrem outlier**, som representerar en tornado

- På föreläsning 3 tog vi upp frågan om **skillnader mellan grupper** i ett datamaterial
- Vi diskuterade hur vi kan bedöma om skillnaderna beror på slumpen eller på att det finns ett mer generellt samband
- Som exempel använde vi Titanics livbåtar, och tittade på hur livbåtsplatser fördelades mellan passagerare i olika klasser
- Till slut ställde vi upp **hypotesen** att livbåtsplatserna var slumpmässigt fördelade och oberoende av biljetklass, och expreimenterade lite
- 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 cirkeldiagram, och jämförde dessa diagram med den verkliga fördelningen
- Den verkliga fördelningen ser inte slumpmässig ut, och vi lutar så åt att det finns ett samband

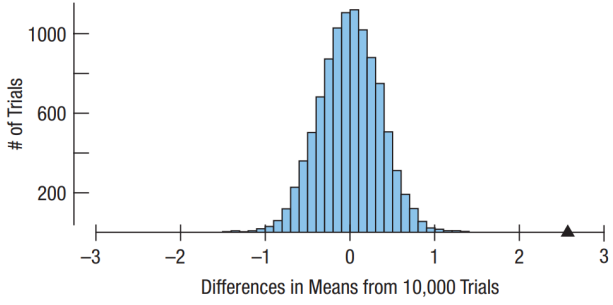
- Vi vill nu göra något liknande, men denna gång vill vi undersöka om **skillnaden mellan två numeriska fördelningar** beror på slumpen eller inte
- Vi antar att vi mäter hastigheten för bilar som kör längs en given gata, och fokuserar särskilt på bilar av märkena BMW och Toyota
- Vi kan sedan illustrera de två hastighetsfördelningarna (en för BMW och en för Toyota) med två låddiagram



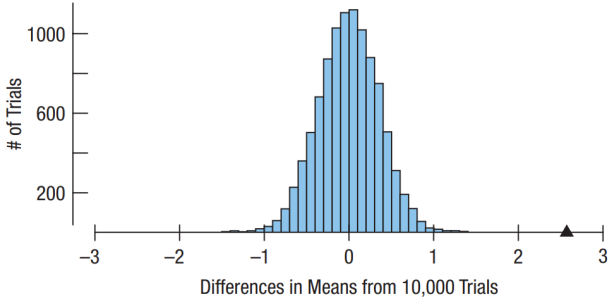
- Medelhastigheten hos de bilar som kör BMW är 2.53 mph högre än medelhastigheten hos de som kör Toyota
- Betyder det att BMW-förare generellt kör snabbare än Toyota-förare, eller beror skillnaden i på slumpen?
- Vi vill veta om skillnaden i medehastighet beror på på att de som kör det ena bilmärket generellt kör snabbare, eller skillnaden beror på slumpen
- Vi kan ställa upp **hypotesen** att hastigheten är oberoende av bilmärke
- Om hypotesen är sann var det enbart slumpen som gjorde så att BMW-förarna körde snabbare i just de fall vi mätte upp

- **Vi gör ett tankeexperiment:** Antag att vi inte känner till vilket bilmärke som är kopplat till vilken hastighet i vår data
- Istället för att dela upp bilarna efter märke, delar vi slumpvis in dem i två grupper som vi kallar A och B
- Eftersom indelningen är slumpmässig, kommer hastigheten nu att vara oberoende av vilken grupp bilarna hamnar i – alltså blir det som att hypotesen är sann
- Vi kan nu beräkna medelhastigheten i de två slumpmässiga grupperna, och notera skillnaden i medelhastigheter
- Vi får nu en indikation om vad vi kan förvänta oss om hypotesen är sann, och det inte finns ett samband mellan bilmärke och medelhastighet

- Vi kan upprepa experimentet från föregående slide **ett stort antal** gånger, för att ta hänsyn till det faktum att slumpen ger viss variation
- Figur 4.5 i De Veaux et al. (2021) visar utfallet av 10,000 sådana experiment



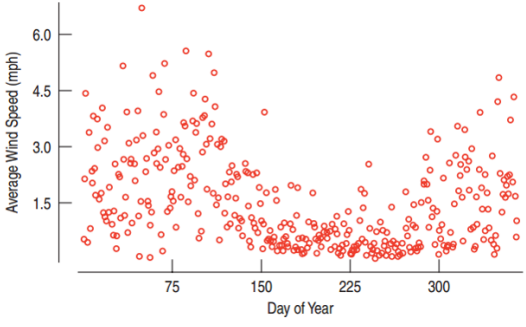
- I studien uppmättes en skillnad (mellan BMW och Toyota) om ca 2.53 mph, denna skillnad är markerad med en triangel i figuren
- Vilka slutsatser kan vi dra?



- Triangeln ligger långt till höger om vad som rimligtvis skulle kunna förväntas om skillnaden bara berodde på slumpen
- Skillnaden är mycket större än vad som vore rimligt om slumpen styrde
- Vi lutar alltså åt att BMW-förare faktiskt kör snabbare

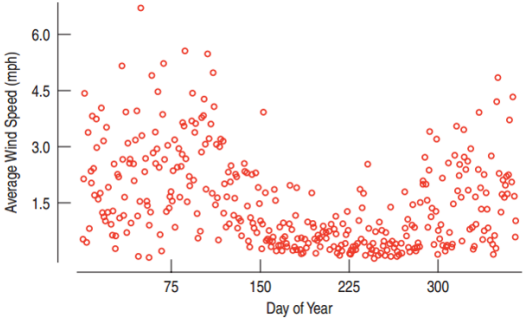
Spridningsdiagram (scatter plot)

- 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, till exempel ett **spridningsdiagram**(*en: scatter plot*)
- Figur 4.6 i De Veaux et al. (2021) visar medelvindstyrkan för varje dag 2011



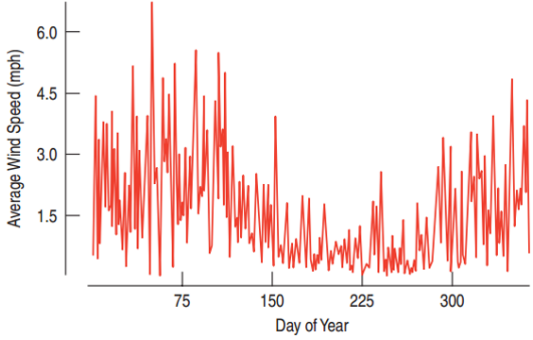
Tidsserier

Data som sträcker sig över tid



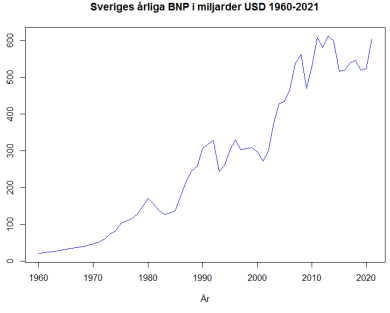
- Spridningsdiagrammet från tidigare är intressant av annan anledning
- På **y-axeln** ser vi medelvindstyrkan förr en given dag
- På **x-axeln** ser vi hur många dagar in på året vi är
- Observationerna är alltså ordnade i tidsordning från vänster till höger, och därmed illustrerar diagrammet en **tidsserie**

- Vi binder gärna samman punkterna för att lättare urskilja mönster i tidsserien



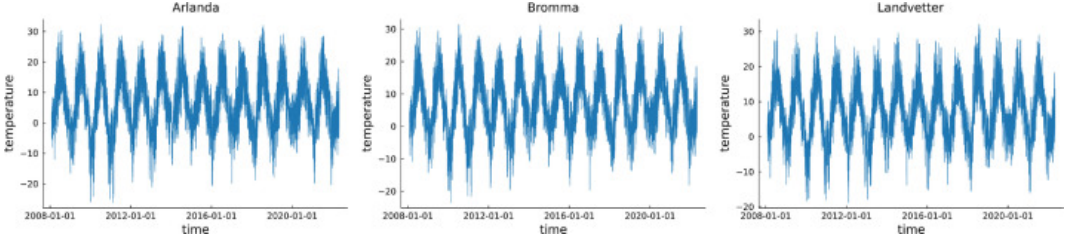
- Vanligtvis är vi intresserade av att titta efter
  - Trender
  - Säsongsvariation
- Trender och säsongsvariationer kan vara viktiga om du vill göra prognoser

- **En trend** är en kontinuerlig förändring som sker över (lång) tid, det kan vara sammanhängande och långsiktiga upp- eller nedgångar
- Figuren visar Sveriges BNP från 1960 till 2020 – finns det trend, och i så fall vilken typ?



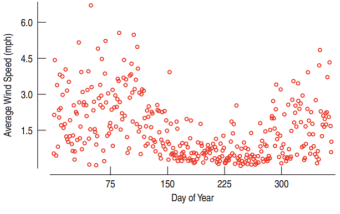
- Även om BNP sjunker vissa år är den övergripande rörelsen växande – om vi drog en rak linje från 1960 till 2020 skulle linjen peka kraftigt uppåt, och vi har alltså **positiv trend**

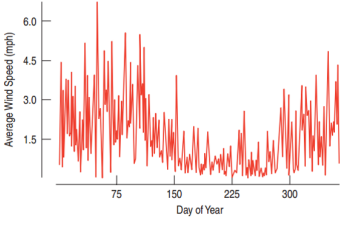
- **Säsongsvariation** är ett mönster som upprepar sig över tydligt avgränsade tidsperioder (t.ex. veckovis, månadsvis årsvis, etc)
- Säsongsvariation syns ofta i exempelvis tidsserier över väder och försäljning (i regel högre temperaturer på sommaren och lägre på vintern, etc)
- Bilden ned visar temperaturer insamlade mellan 1 februari 2008 och 1 maj 2022 vid tre svenska flygplatser – tagen från Villani et al. (2022)

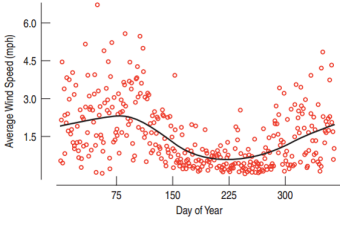


Utjämning (smoothing)

- Bilderna till vänster och i mitten är båda kaotiska
- Det högra diagrammet innehåller en utjämnad kurva som visar seriens stora rörelser
- En enkel metod för utjämning kallas för **glidande medelvärde** (*en: moving average*)
- För varje tidpunkt beräknar vi ett medelvärde av de punkter som ligger närmast i tid



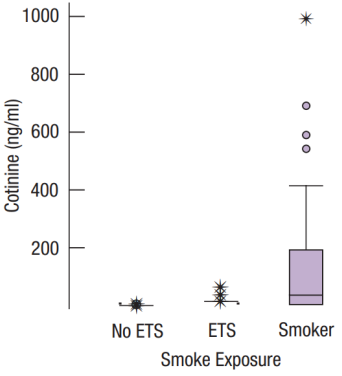




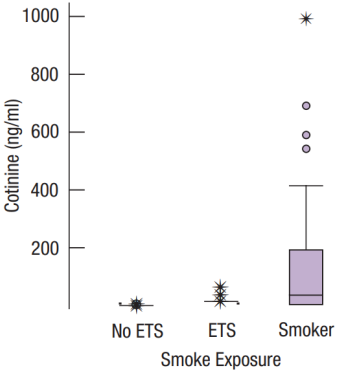
Transformationer

Varför transformera?

- Vi motiverar **transformation** av variabler med hjälp av en forskningsstudie
- Studien undersöker om exponering för rökning påverkade nivån av kotinin i blodet (nedbrytningsprodukt av nikotin)
- Deltagarna i studien delades in i tre grupper:
  - Rökare (Smoker)
  - Passiva rökare (ETS, *exposed to smoke*)
  - De som inte utsattes för någon rök (No ETS)
- Vi har alltså en kategorisk variabel (grupptillhörighet) och en numerisk variabel (mängd kotonin i blodet, nanogram/ml)

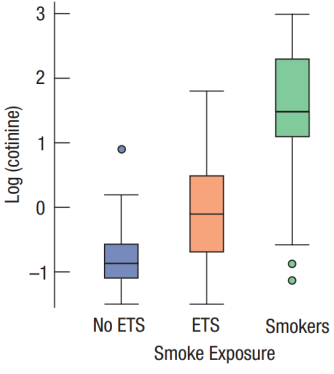


- I figuren finns tre låddiagram, där varje diagram visar fördelningen av kotonin-nivå i respektive grupp
- Ser ni några problem med diagrammet?



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

- Med hjälp av en transformation kan vi göra diagrammen mer lättläsliga, i detta fall kan vi transformera intressevariabeln från kotinin till log(kotinin)



- Vi ser nu tydligt att de som exponerats för rök har högre halter av kotinin
- Men: nu har vi **log(kotinin)** på y-axeln, och **inte** mängden kotinin

- 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.

- Antag att vi vet att $\log(a) = 1.2$ – kan vi då säga något om värdet på a själv?
- Ja, vi kan använda den andra regeln från föregående slide, nämligen

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

- Vi kan stoppa in a istället för y, och få

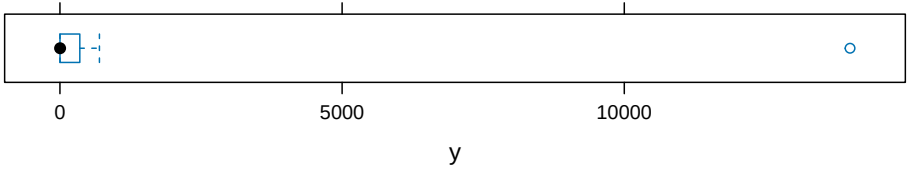
$$a = e^{\log(a)}$$

- Eftersom $\log(a) = 1.2$ är given i "uppgiften" har vi så att

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

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

```
> library(mosaic) # Behövs för bwplot()
> 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, då nästan alla observationer är ihoptryckta i diagrammets vänstra del

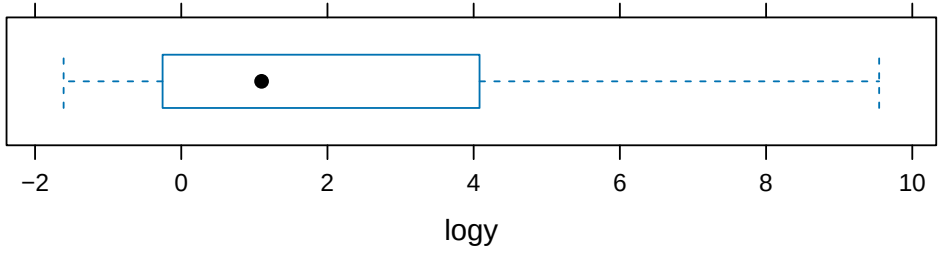
- Vi skapar nu variabeln $\log y$, som är logaritmen av y, och skriver ut den nya variabelns värden avrundade till 3 decimaler

```
> logy <- log(y) # Skapa logaritmerat
>
> print(round(logy, 3)) # Skriv ut avrundad logaritmerad vektor
> [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

- 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

- Vi kan transformera tillbaka våra värden till orginalskalan med formeln

$$y = e^{\log(y)}, \quad \text{där} \quad e \approx 2.718$$

```
> 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
- Notera att e^x i R skrivs `exp(x)`

Transformationer

- Logaritmering är användbart i otroligt många statistiska tillämpningar
- Men det 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!
- På den här kursen kommer transformationer inte handla om matematik, utan mer om att pröva sig fram

Tack för idag!!