Stockholm University

Föreläsning 5: Jämföra fördelningar. Tidsserier. Transformationer

Matias Quiroz[1]

[1]Statistiska institutionen, Stockholms universitet

VT 2023

Innehåll

- Undersöka samband mellan en numerisk variabel och en kategoriska variabel.
- Låddiagram.
- Tidsserier.
- Transformationer av data.

Jämföra fördelningar för två kategoriska variabler

- Under förra föreläsningen lärde vi oss att studera samband mellan två kategoriska variabler.
- Vi åstad kom detta genom att först tag fram den betingade fördelningen för en variabel, dvs en fördelning för en variabel givet ett värde för den andra.
- Exemplet med husdjurspreferens betingat på kön:

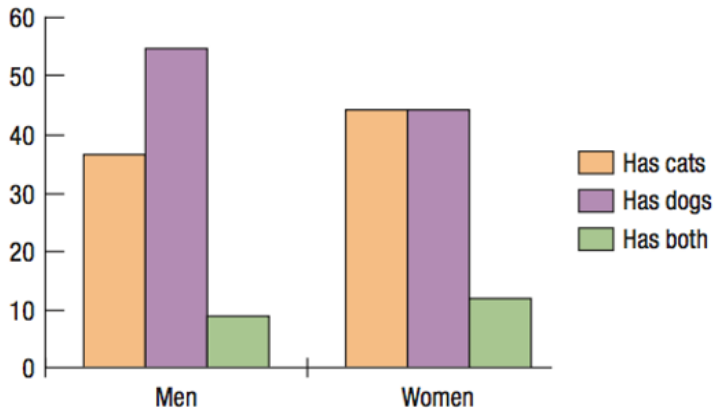


Figure 1: Figur 3.2 i De Veaux et al. (2021).

- Notera att vi jämför (betingade) fördelningar mellan två grupper (män och kvinnor)!
- Andra grafiska verktyg: staplade stapeldiagram, mosaic plot.

Jfr fördeln. mellan grupper för en numerisk variabel, forts.

- Ett vanligt förekommande fall är att man studera **sambandet mellan en numerisk variabel och en kategorisk variabel**.
- Precis som förut så jämför vi födelningarna mellan två (eller flera) grupper.
- Grupperna motsvarar, precis som förut, de olika värden av den kategoriska variabeln som vi betingar på.
- Skillnaden är att variabeln vi gör fördelningen för är numerisk!
- Föreläsning 3: Visualiseringsverktyg för numeriska variablers fördelningar
  1. Histogram.
  2. Täthetsplot.
  3. Stam och blad diagram.
  4. Punktplot.
- Gör en valfri bland 1.–4. för varje grupp för att jämföra fördelningar.

Jfr fördeln. mellan grupper för en numerisk variabel, forts.

- Dagliga medelvindhastigheter under 2011 i western Massachusetts:

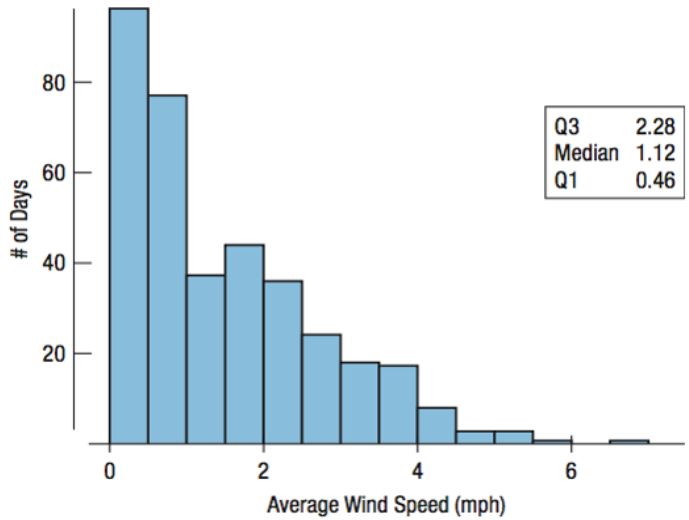


Figure 2: Figur 4.1 i De Veaux et al. (2021).

- Kommentarer:
  - Hälften av observationerna < 1.12 mph.
  - En extrem observation 6.73 mph.
  - Fördelningen är skev åt höger (längre svans åt höger).
- Datasetet innehåller också datum för varje observation.

Jfr fördeln. mellan grupper för en numerisk variabel, forts.

- Om vi skapar en kategorisk variabel season (säsong), med två utfall `Spring/Summer` och `Fall/Winter` får vi följande fördelningar:

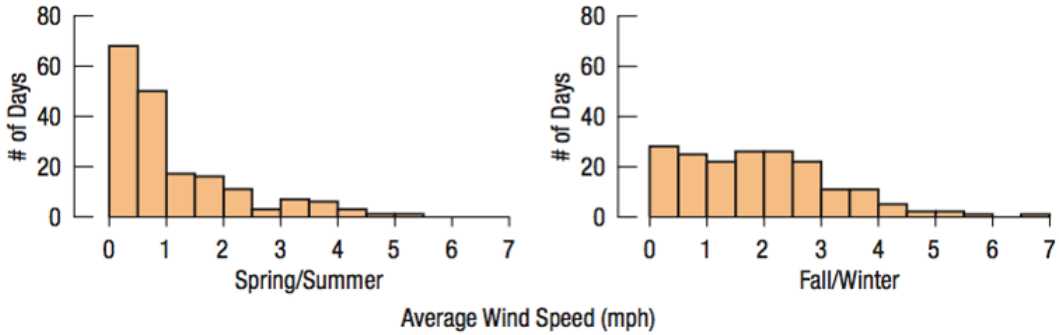


Figure 3: Figur 4.2 i De Veaux et al. (2021).

- Kommentarer:
  - Fördelningen under `Spring/Summer` är skev åt höger (längre svans åt höger).
  - Fördelningen under `Fall/Winter` är mer likformig.
  - De flesta dagar under `Spring/Summer` blåser det < 1 mph.
- Vi kan se dessa som betingade fördelningar (betingat på `season`).
- Vilken är den marginella fördelningen för medelvindhastigheter?

Jfr fördeln. mellan grupper för en numerisk variabel, forts.

- Några beskrivande mått för fördelningarna:

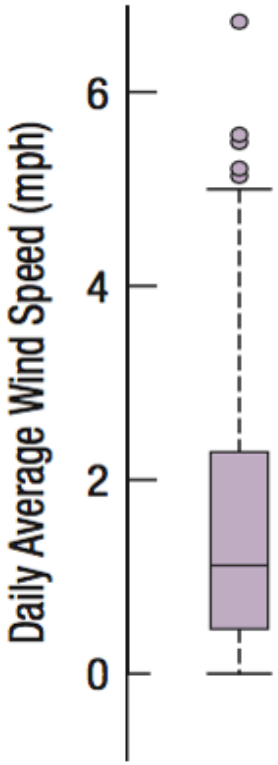
Summaries for *Average Wind Speed* by Season

Season	Mean	StdDev	Median	IQR
Summer	1.11	1.10	0.71	1.27
Winter	1.90	1.29	1.72	1.82

Figure 4: Tabell från s.123 i De Veaux et al. (2021).

- Dessa styrker vårt iaktagande att det blåser mer under vinterhalvåret.
- Antag att vi istället för `season` variabeln har en månadsvariabel (också kategorisk).
- Det blir svåröverskådligt att rita 12 histogram.
- Ett låddiagram (**box plot** på engelska) passar bättre i detta fall. Kallas ibland för lådagram.
- Ett låddiagram visar **median, IQR och outliers tydligare** än ett histogram.

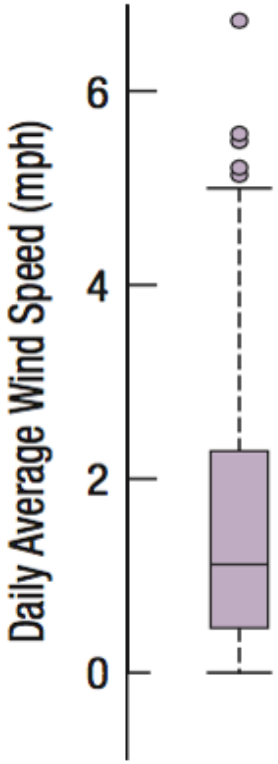
Jfr fördeln. mellan grupper för en numerisk variabel, forts.



- Höjden på toppen av lådan är Q3.
- Höjden på botten av lådan är Q1.
- Lådan har höjd IQR = Q3-Q1.
- Horisontella strecket i lådan är Q2 (medianen).
- Låt $Q3 + 1.5 \cdot IQR$ och $Q1 - 1.5 \cdot IQR$ vara osynliga horisontella streck.
- Observationer utanför de osynliga strecken räknas som outliers och ritas in som cirklar.
- Största observationen innanför de osynliga strecken bildar det övre horisontella strecket utanför lådan.
- Minsta observationen innanför de osynliga strecken bildar det undre horisontella strecket utanför lådan.
- Observationer som hamnar $3 \cdot IQR$ över Q3 eller under Q1 ritas in med en annan symbol.

Jfr fördeln. mellan grupper för en numerisk variabel, forts.

- Information att utläsa i ett låddiagram.



  - Lådan visar var den mittersta 50% av data ligger.
  - Om medianen är centrerad i lådan så är mittersta 50% av data symmetrisk.
  - Om medianen inte är centrerad i lådan så är mittersta 50% av data skev.
  - Om de streckade vertikala linjerna inte är lika långa är datan skev.
  - Om många observationer hamnar som outliers bör man vara uppmärksam.

- Låddiagrammet syftar inte till att uppmana att ta bort outliers, utan endast uppmärksamma att de finns.

Jfr fördeln. mellan grupper för en numerisk variabel, forts.

- Fördelning över dagliga medelvindhastigheter för varje månad:

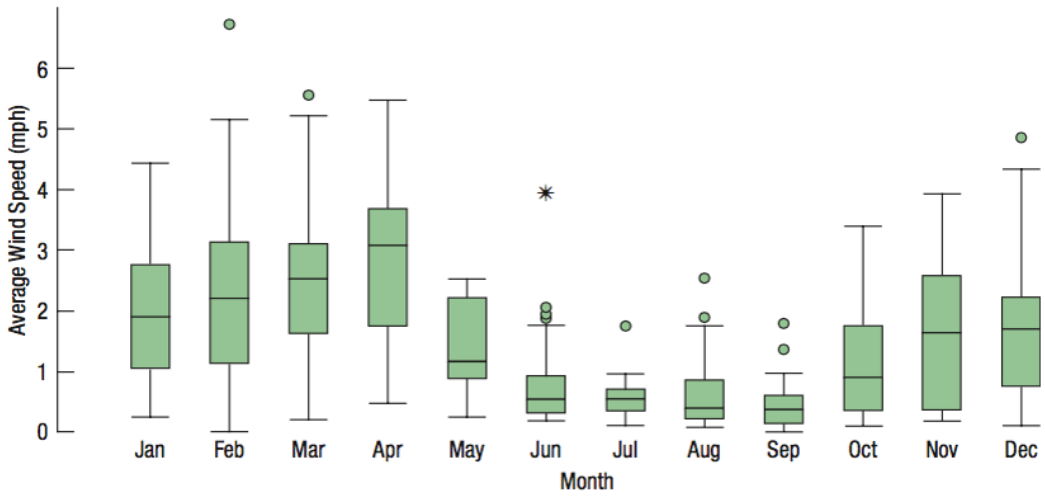


Figure 5: Figur 4.3 i De Veaux et al. (2021).

- Kommentarer:
  - Under sommarmånaderna blåser det mindre.
  - Fördelningen varierar mindre sommarmånaderna.
  - Det blåser betydligt mer under Oktober-April.
  - Fördelningen varierar mer under Oktober-April.
  - Extrem outlier under juni månad. Visade sig vara en ovanlig tromb (tornado).
- Mycket mer överskådligt och informativt än 12 histogram!

Jfr fördeln. mellan grupper för en numerisk variabel, forts.

- På en väg med hastighetsgräns 20 mph uppmättes hastigheten för 500 bilar samt i vilken färdriktning (uppåtriktning eller nedåtriktning) varje bil körde. Man ville utreda om bilar körde snabbare när de färdades i uppåtriktning.
- Följande låddiagram visar hastigheterna för de två olika grupperna:

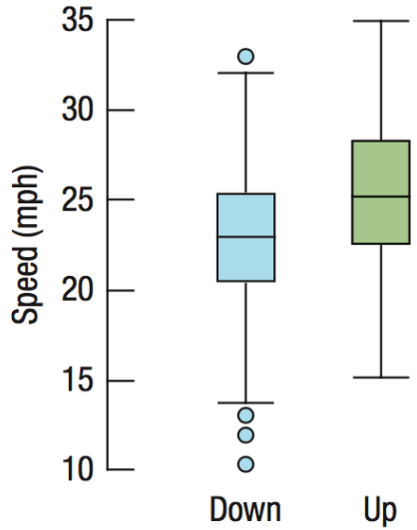


Figure 6: Figur 4.4 i De Veaux et al. (2021).

- Skillnaden mellan medelvärdena (uppåtrikning–nedåtriktning) var 2.53 mph.
- Kan skillnaden bero på slumpen och inte i vilken riktning bilarna färdades (dvs inget samband mellan hastighet och riktning)?

Vad händer om vi får nya data?

- Tankeexperiment: Antag att vi slumpar ut riktningar för varje bil.
- Då skulle bilens hastighet vara oberoende av dess riktning.
- Vi kan då räkna ut ett nytt medelvärde för varje grupp och beräkna en ny skillnad mellan medelvärdena (uppåtrikning–nedåtriktning).
- Om vi upprepar detta 10000 gånger:

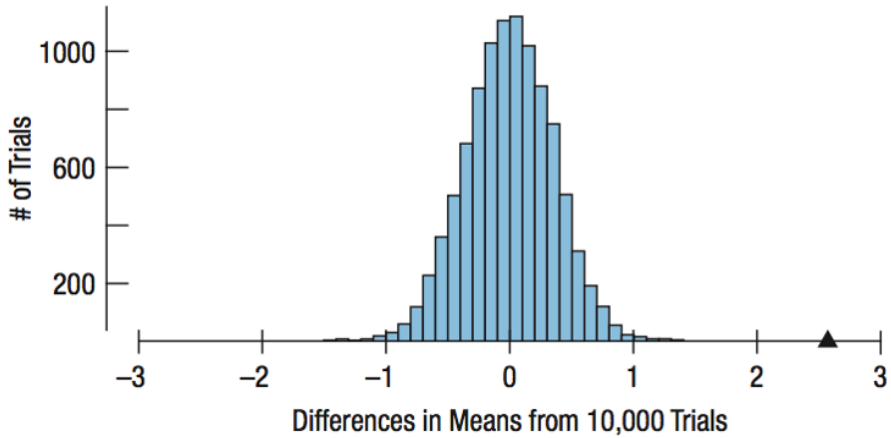


Figure 7: Figur 4.5 i De Veaux et al. (2021).

- Vilken slutsats ger tankeexperimentet om **hypotesen** att den färdriktningen bilen färdas i inte har något samband med hastigheten?

Vad händer om vi får nya data?, forts.

- Betrakta de simulerade värdena igen. Notera att fördelningen är symmetrisk:

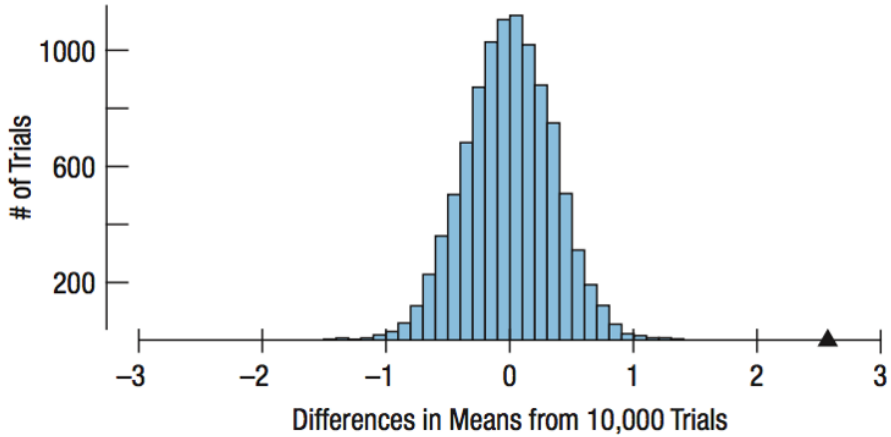


Figure 8: Figur 4.5 i De Veaux et al. (2021).

- Denna fördelning kallas för samplingfördelningen (**sampling distribution** på engelska) för skillnaden mellan medelvärden för de olika grupperna.
- En samplingfördelning är kanske **det viktigaste begreppet i kursen Statistik och Dataanalys I**.
- En samplingfördelning beskriver hur en viss **kvantitet varierar från stickprov till stickprov**.
- Den specifika kvantiteten här är skillnaden mellan medelvärden för de olika grupperna. Kan vara andra kvantiteter också.

Tidsserier

- Dagliga medelvindhastigheter under 2011 i western Massachusetts kan betraktas som en tidsserie (**time series** på engelska).
- En tidsserie kännetecknas av att det finns en naturlig **tidsordning i observationerna**. Naturligt att plotta varje observation mot sitt tidsindex.
- Exempel på ett punktdiagram (**scatter plot** på engelska) för en tidsserie:

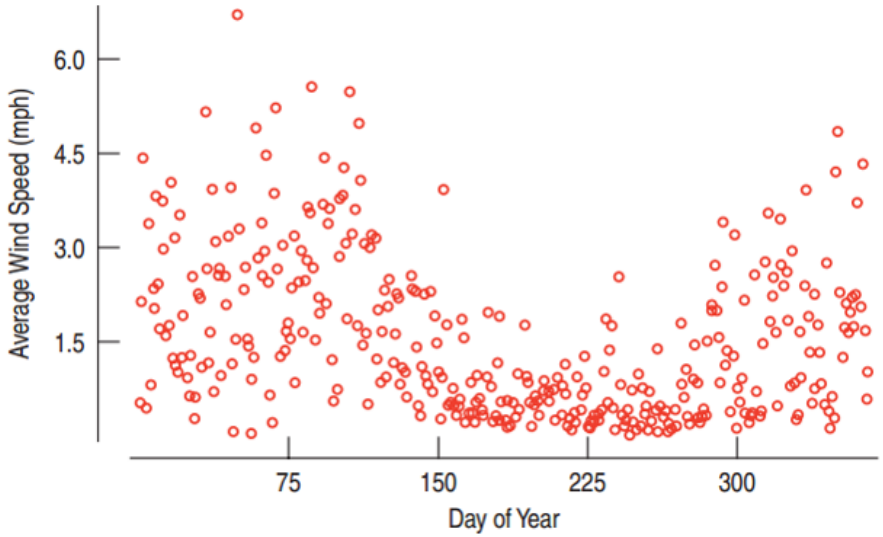


Figure 9: Figur 4.6 i De Veaux et al. (2021).

- Låddiagrammen grupperade ihop observationerna i månader. Här ser vi istället varje observation för sig.

Tidsserier, forts.

- Eftersom observationerna inte är grupperade så är tid en numerisk variabel. Mer om samband mellan två numeriska variabler i Föreläsning 7.
- En tidsserieplot binder ihop punkterna med en linje:

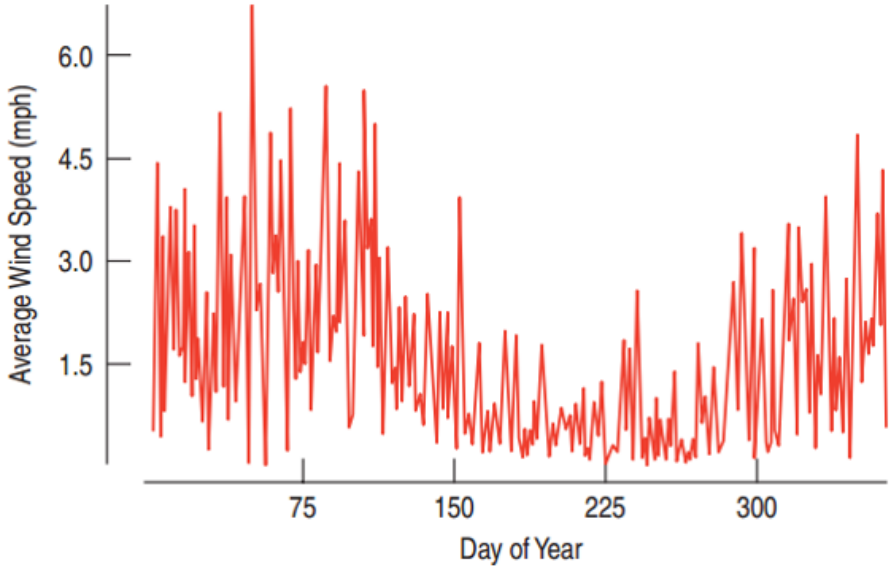


Figure 10: Figur 4.7 i De Veaux et al. (2021).

- Kan ibland underlätta med linjer för att tydligare se mönster i tidsserien.
- Vilka mönster letar man efter?
  1. Finns det en **trend** i tidsserien?
  2. Finns det **säsongsvariation**?

Tidsserier, forts.

- Exempel på en tidsserie som har en (positiv) trend är Sveriges bruttonational produkt (BNP).

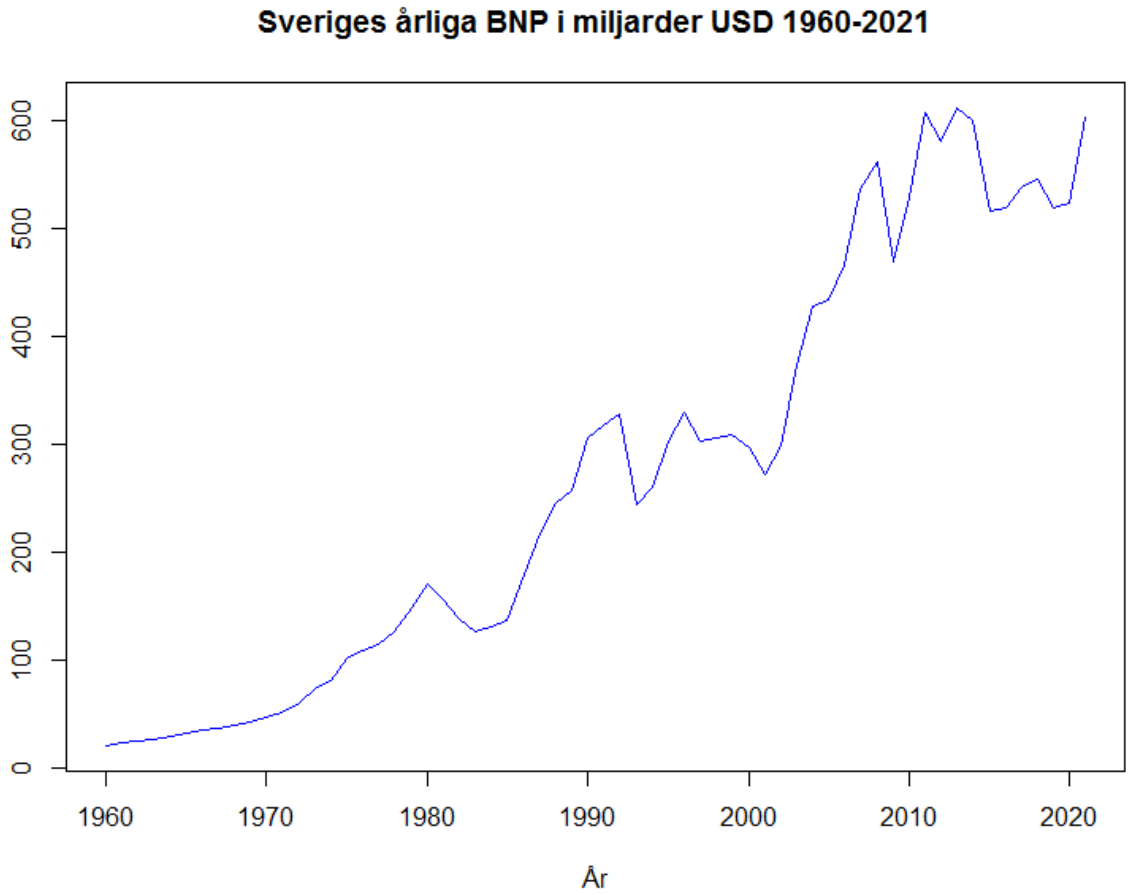


Figure 11: Data från `datacommons.org`.

Tidsserier, forts.

- Exempel på tidsserier som har säsongsvariation är temperaturer.
- Temperaturer samlade på timfrekvens mellan 1 februari 2008 till 1 maj 2022 vid tre svenska flygplatser.

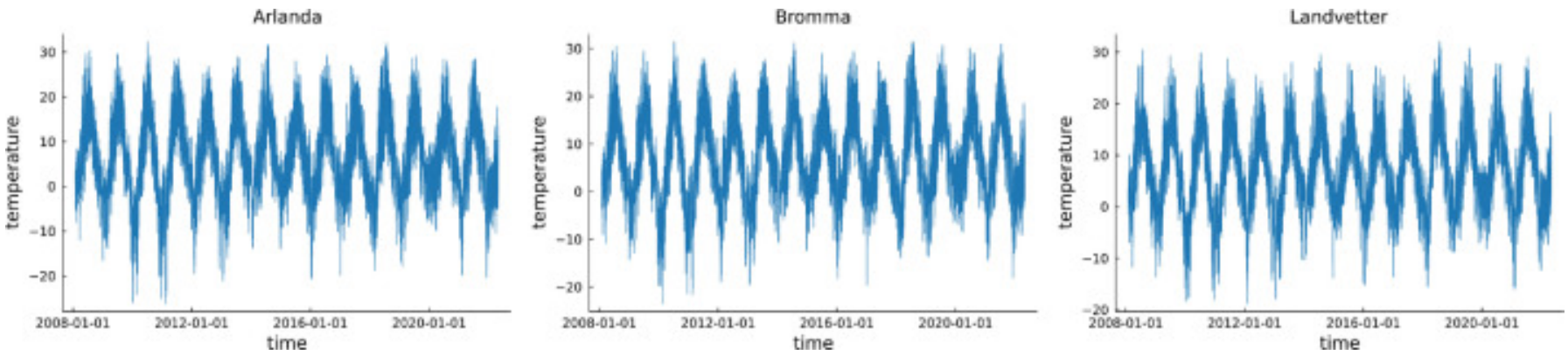


Figure 12: Figur från Villani et al. (2022).

Tidsserier, forts.

- För att tydligare se mönster i tidsserien kan man använda utjämningsmetoder (**smoothing methods** på engelska).
- Förenklat uttryckt så beräknar en utjämningsmetod ett viktat medelvärde av en tidsserie genom att ge observationer nära i tiden en större vikt.
- Exempel på utjämning av dagliga medelvindhastigheter:

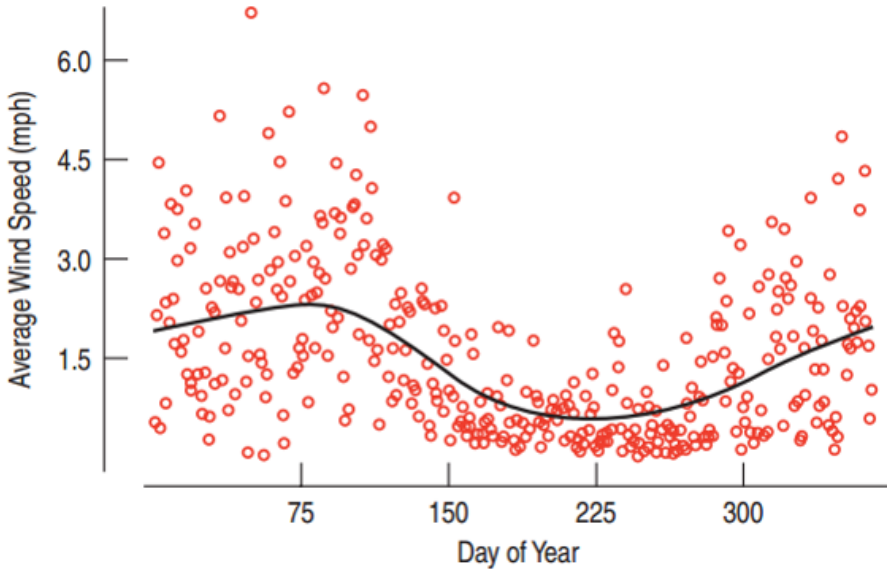


Figure 13: Figur 4.8 i De Veaux et al. (2021).

- Utämningsmetoden ovan finns implementerad i R-funktionen `loess`.

Transformationer

- Deskriptiv statistik har många verktyg för att illustrera fördelningen av en variabel.
- För en numerisk variabel är histogrammet eller en täthetsplot de mest användbara.
- Statistik handlar om att modellera data (beskriva data) med hjälp av **teoretiska fördelningar**.
- Histogrammet och täthetsplotten är en **empiriska versioner** av den teoretiska fördelningen.
- **Normalfördelningen** (nästa föreläsning) blir den första teoretiska fördelningen vi stöter på.
- Normalfördelningen är en **symmetrisk fördelning**.
- Skeva data kan ofta **transformeras för att få en mer symmetrisk fördelning**. Normalfördelningsantagandet blir då rimligare.

Transformationer, forts.

- Andra anledningar för att vilja transformera data:
  - Svårt att avgöra om icke-transformerade värden är outliers (nästa föreläsning ger en regel som bygger på normalfördelningen).
  - Svårt att sammanfatta fördelningen med centrala värden (medelvärde, median) och spridningsmått (standardavvikelse, IQR).
  - Figurer med icke-transformerade data kan vara svåra att tyda.
  - Svårare att se samband med icke-transformerade värden.

Transformationer, forts.

- Verkställande direktörers kompensation för Forbes 500 företag:

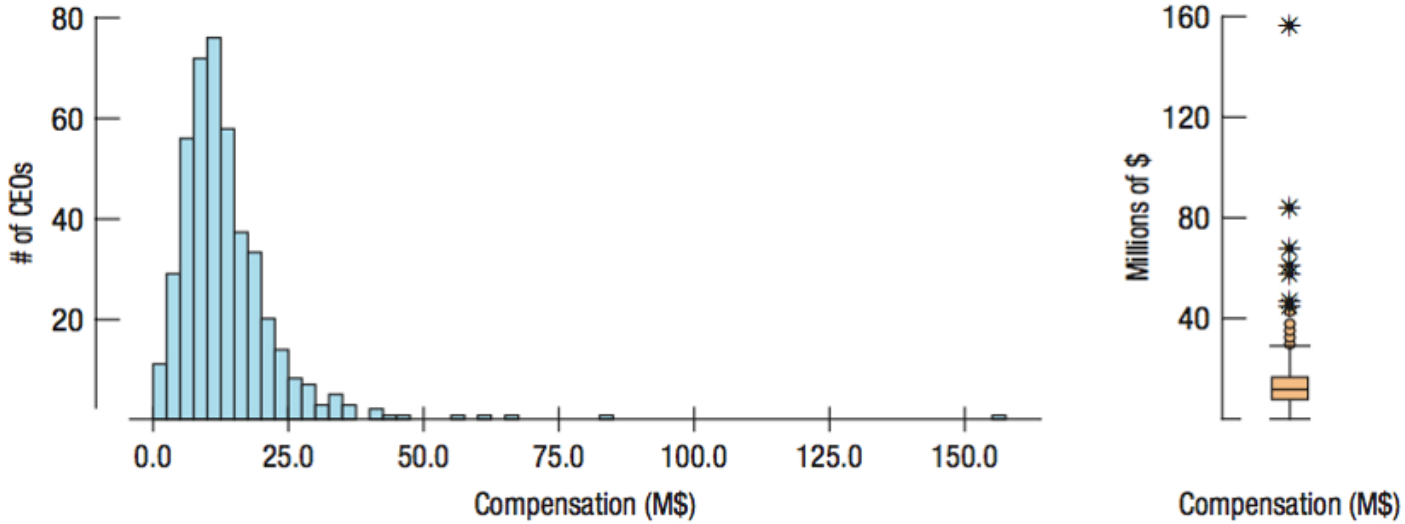


Figure 14: Figur 4.9 i De Veaux et al. (2021).

- Kommentarer:
  - Fördelningen är skev åt höger (lång svans till höger).
  - Låddiagrammet visar 7 extremvärden.
  - Stor skillnad på medianen ($11 841 179) och medelvärdet ($13 903 006).

Transformationer, forts.

- En logaritmisk transformation med basen 10, dvs $\log_{10}(y)$ ger:

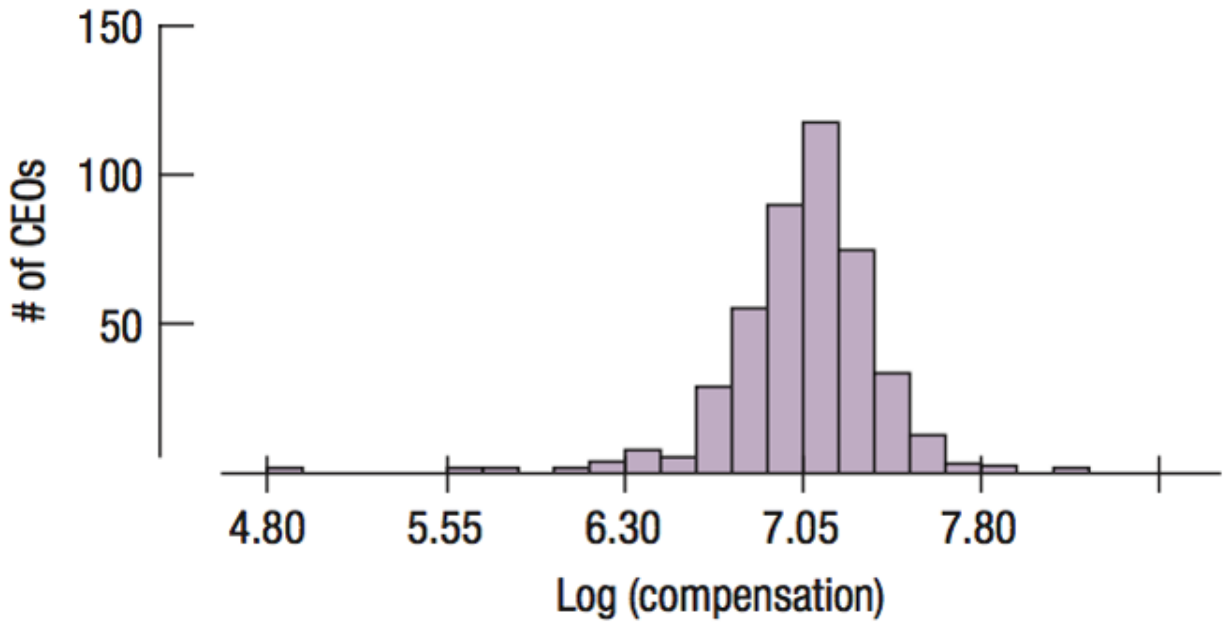


Figure 15: Figur 4.10 i De Veaux et al. (2021).

- Kommentarer:
  - Fördelningen är mycket mer symmetrisk.
  - 4 extremvärden (enligt låddiagrammets definition).
  - Ingen stor skillnad på medianen och medelvärdet, båda runt 7.

Transformationer, forts.

- Vilken transformation ska man göra?
- För variabler vars fördelningar är skeva åt höger: $\log_{10}(y)$, $\ln(y)$, eller $1/y$.
- ln är den så kallade **naturliga logaritmen**. En logaritm med naturliga talet $e \approx 2.71828$ som bas.
- Ofta används log istället för notationen ln. Boken använder notationen log istället för $\log_{10}$. Viktigt att kolla vilken logaritm som används!
- För variabler vars fördelningar är skeva åt vänster: y^2.
- Att transformera är mer av en konst än vetenskap. Testa er fram!
- Tänk på att logaritmen ändrar skalan på mätningarna! I VD exempel var medianlönen efter transformation runt 7 log-dollar (som motsvarar 10^7 dollar)

Transformationer, forts.

- En studie undersökte om exponering för rökning påverkade kotininnivån (nedbrytningsprodukt av nikotin) i blodet.
- Exponering för rökning delades in i tre utfall: rökare, passiva rökare (ETS) och de som inte utsattes för någon rök (No ETS).
- ETS står för exposed to smoke.

Transformationer, forts.

- Låddiagrammen för kotininnivån mot rökexponering.

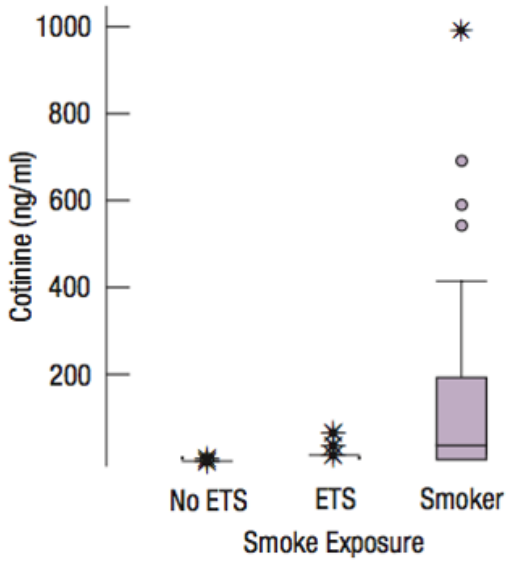


Figure 16: Figur 4.11 i De Veaux et al. (2021).

- Kommentarer:
  - Låddiagrammen för passiva rökare och de som inte utsatts för rök oläsliga.
  - Många outliers i de två första låddiagrammen.
  - Svårt att se om det föreligger en skillnad mellan de passiva rökarna och de som inte utsatts för rök.

Transformationer, forts.

- Låddiagrammen för log-kotininnivån mot rökexponering.

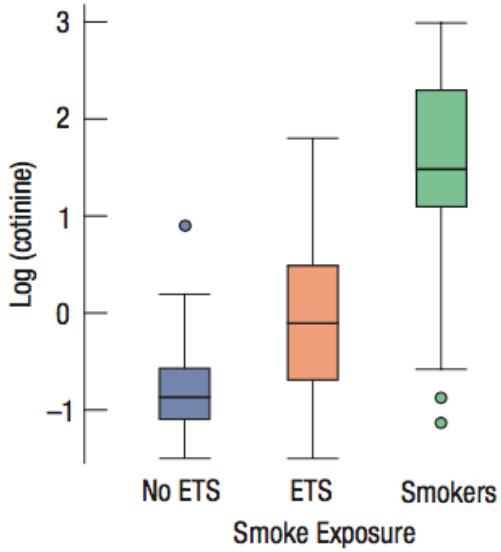


Figure 17: Figur 4.12 i De Veaux et al. (2021).

- Kommentarer:
  - Mycket enklare att utläsa.
  - Inga outliers i log-skala.
  - De passiva rökarna verkar ha högre log-kotininnivå än de som inte utsatts för rök.

References I

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

Villani, M., Quiroz, M., Kohn, R., and Salomone, R. (2022). Spectral subsampling MCMC for stationary multivariate time series with applications to vector ARTFIMA processes. *Econometrics and Statistics*, (In Press).