Stockholm University

Föreläsning 6: Standardisering och normalfördelningen

Matias Quiroz[1]

[1]Statistiska institutionen, Stockholms universitet

VT 2023

Innehåll

- Standardavvikelse som jämförelsemått.
- Skiftning och skalning av data.
- Standardisering och z-värden.
- Normalfördelningen.
- Population kontra stickprov.
- Normalfördelningstabell.
- Normalfördelningsplot.

Standardavvikelse som jämförelsemått

- Ibland vill vi jämföra observationer från två olika fördelningar och bedömma vilken som är mest avvikande från sin fördelning (grupp).
- Exempel: Under damernas sjukamp i OS 2016 vann:
  - Nafissatou Thiam längdhoppstävlingen med ett längdhopp på 6.58 meter, 0.41 meter längre än genomsnittet.
  - Katarina Johnson-Thompson 200 meter med en tid på 23.26 sekunder, 1.32 sekunder snabbare än genomsnittet.
- Vilket av resultaten är mest avvikande?
- Notera:
  - Observationerna har olika enheter (meter och sekunder).
  - Grupperna har olika nivåer (medelvärden runt 6 och runt 25).
  - För att uttala oss om avvikande i förhållande till de olika grupperna räcker inte genomsnittet. **Vi måste ha en uppfattning om spridningen inom grupperna**.
- Vårt mål är att få ett mått på avvikelse som inte beror på varken enheter eller vilka medelvärden grupperna har.

Standardavvikelse som jämförelsemått, forts

- Medelvärdet $\overline{y}$ och standardavvikelsen (SD) s defineras som

$$\overline{y} = \frac{\sum y}{n} \quad \text{och} \quad s = \sqrt{\frac{\sum (y - \overline{y})^2}{n-1}}.$$

- Tabell för de olika grenarna:

	Längdhopp (m)	**200 m (s)**
Medelvärde	6.17	24.58
Standardavvikelse	0.247	0.654
Individuell obs	6.58	23.26

- För längdhopp (notera: större är bättre):
  - 1 SD ovanför medelvärdet: $6.17 + 1 \cdot 0.247 = 6.471$.
  - 2 SD ovanför medelvärdet: $6.17 + 2 \cdot 0.247 = 6.664$.
- Observationen 6.58 är ungefär 1.5 standardavvikelser från medelvärdet (till höger om medelvärdet).

Standardavvikelse som jämförelsemått, forts

- Tabell för de olika grenarna:

	Längdhopp (m)	**200 m (s)**
Medelvärde	6.17	24.58
Standardavvikelse	0.247	0.654
Individuell obs	6.58	23.26

- För 200 m (notera: mindre är bättre):
  - 1 SD under medelvärdet: $24.58 - 1 \cdot 0.654 = 23.926$.
  - 2 SD under medelvärdet: $24.58 - 2 \cdot 0.654 = 23.272$.
- Observationen 23.26 är strax över 2 standardavvikelser från medelvärdet (till vänster om medelvärdet).
- Katarina Johnson-Thompsons resultat mer imponerande (**avviker fler SD från medelvärdet**).

Standardavvikelse som jämförelsemått, forts

- Tabell för de olika grenarna:

	Längdhopp (m)	**200 m (s)**
Medelvärde	6.17	24.58
Standardavvikelse	0.247	0.654
Individuell obs	6.58	23.26

- Låt oss räkna de exakta antal SD varje observation avviker från sitt medelvärde.

- Nafissatou Thiam (längdhopp):

$$6.17 + z \cdot 0.247 = 6.58 \implies z = \frac{6.58 - 6.17}{0.247} \approx 1.659.$$

- Katarina Johnson-Thompson (200 m):

$$24.58 + z \cdot 0.654 = 23.26 \implies z = \frac{23.26 - 24.58}{0.654} \approx -2.018.$$

Standardavvikelse som jämförelsemått, forts

- Antalet standardavvikelser en observation avviker från sitt medelvärde är en sådan viktig kvantitet i statistik att den fått ett eget namn och beteckning.
- Den kallas för z-värde (**z-score** på engelska) och bokstaven z används för att beteckna den.
- Den räknas enligt formeln:

$$z = \frac{y - \overline{y}}{s}.$$

- Enkel tolkning: Antalet standardavvikelser y avviker från sitt medelvärde.
- Om ett z-värde är positivt så ligger observationen y till höger om medelvärdet.
- Om ett z-värde är negativt så ligger observationen y till vänster om medelvärdet.
- Vart ligger observationen y om z-värdet är 0?

Skiftning och skalning av data

- Notera att z-värdet är en transformation av y,

$$z = \frac{y - \overline{y}}{s}.$$

- I ord: Dra av medelvärdet $\overline{y}$ från y och dela med standardavvikelsen för y.
- "Dra av" kallas for skiftning eftersom det skiftar fördelningen för y.
- Exempel: Viktfördelning for 80 män mellan 18–24 år och $\approx$ 172–178 cm:

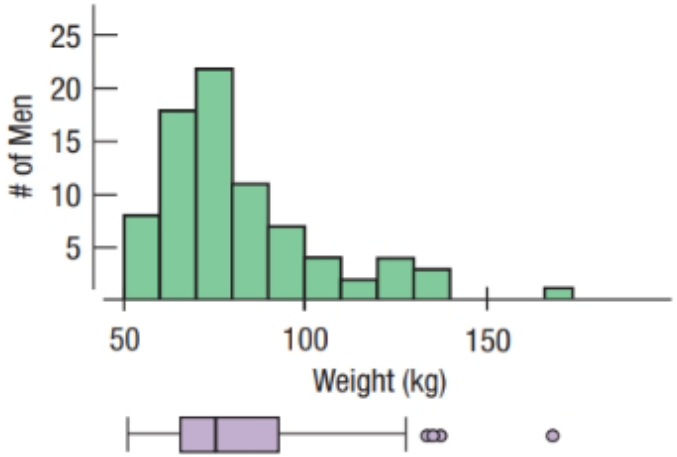


Figure 1: Figur 5.2 i De Veaux et al. (2021).

Skiftning och skalning av data, forts.

- Gruppens medelvärde är 82.36 kg. Rekommenderar maxvikt är 74 kg.
- I genomsnitt har gruppen $82.36 - 74 = 8.36$ kg övervikt.
- Dra av 74 kg för att studera variabeln "kg over rekommenderad vikt":

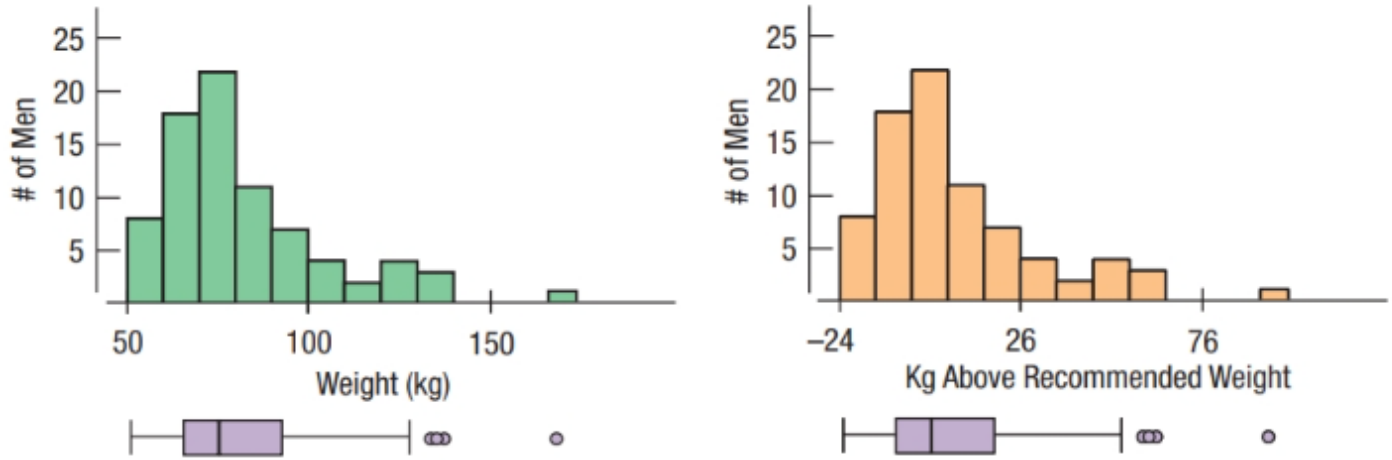


Figure 2: Figur 5.2 och 5.3 i De Veaux et al. (2021). Original (vänster) och skiftad (höger) fördelning.

- Att dra av en konstant a skiftar alla observationer till vänster med a enheter, inklusive medelvärdet.
- Hur har fördelningens spridning påverkats?

Skiftning och skalning av data, forts.

- **Fördelningens spridning påverkas inte om vi adderar eller substraherar en konstant**.
- Detta betyder att variationsbredden, IQR och standardavvikelsen för skiftade data är samma som för oskiftade.
- Hur påverkas en fördelnings form, medelvärde, och spridning, om vi väljer att mäta i en annan skala?
- Antag att vi istället mäter i pounds (1 kg $\approx$ 2.2 pounds)

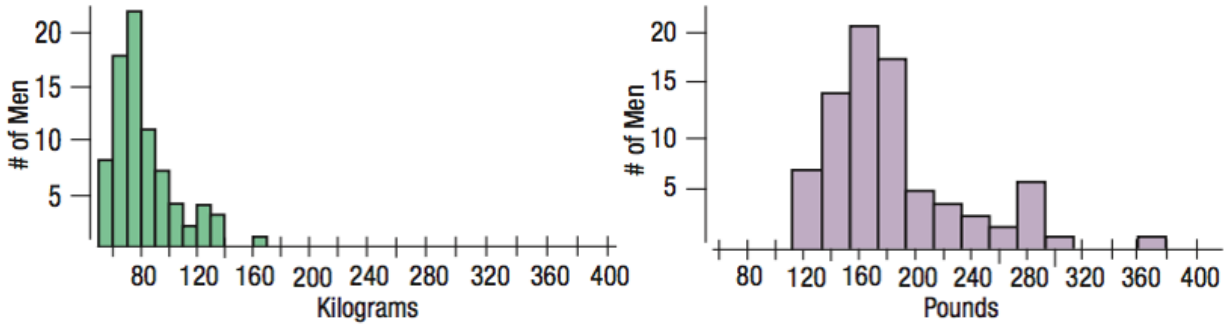


Figure 3: Figur 5.4 i De Veaux et al. (2021). Original (vänster) och multiplicerad med 2.2 (höger) fördelning.

- Fördelningens är forfarande unimodal och skev åt höger.

Skiftning och skalning av data, forts.

- Medelvärdet har multiplicerats med 2.2:

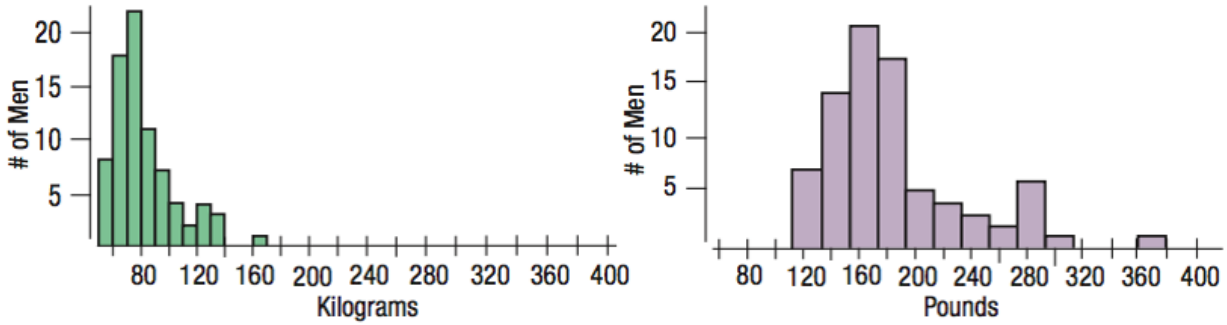
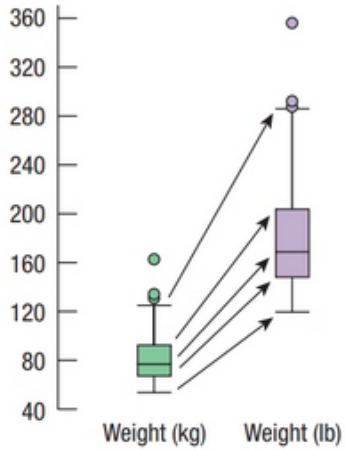


Figure 4: Figur 5.4 i De Veaux et al. (2021). Original (vänster) och multiplicerad med 2.2 (höger) fördelning.

- Fördelningens spridningsmått har multiplicerats med 2.2:



	Weight (kg)	Weight (lb)
Min	54.3	119.46
Q1	67.3	148.06
Median	76.85	169.07
Q3	92.3	203.06
Max	161.50	355.30
IQR	25	55
SD	22.27	48.99

Figure 5: Figur 5.5 i De Veaux et al. (2021).

Skiftning och skalning av data, forts.

- När vi multiplicerar eller dividerar med en konstant så påverkas:
  - Alla lägesmått (medelvärde och medianvärde).
  - Alla spridningsmått (variationsbredd, Q1, Q3, IQR, standardavvikelse).
- **Alla dessa mått multipliceras (eller divideras) med samma konstant som vi multiplicerar (eller dividerar) data med**.
- Vi har visat detta heuristiskt för ett specifikt exempel. Kan även visas matematiskt utifrån definitionerna och då gäller resultaten generellt.
- Ett mycket viktigt resultat är medelvärdet och standardavvikelsen för z-värdena.

Skiftning och skalning av data, forts.

- Definitionen av z-värdet är
$$z = \frac{y - \overline{y}}{s}.$$
- $y - \overline{y}$ har medelvärde 0 och standardavvikelse s.
- $z = (y - \overline{y})/s$ har **medelvärde 0 och standardavvikelse 1** ($s/s = 1$).
- Vi säger att z-värdet standardiserar y (**standardize** på engelska).
- Standardisering (**standardizing** på engelska) centrerar fördelningens för y den kring 0 och skalar om dess spridning så standardavvikelsen är 1.
- Vad är variansen av de standardiserade z-värdena?
- Det typiska värdet för z är 0.
- Vad anses vara ett stort z värde? Är $z = 1$ stort? Eller $z = 2$?
- Detta beror naturligtvis på hur fördelningen för z ser ut.

Normalfördelningen

- Vi är redo att presentera vår första teoretiska fördelning!
- För att förstå vad en teoretisk fördelning är på ett djupare plan så behöver vi **sannolikhetsteori**.
- Sannolikhetsteori gås igenom på andra delen av kursen.
- Teaser på vad som väntar: Villanis `widgets`.
- I den här delen av kursen använder vi normalfördelningen praktiskt utan en djupare förståelse.
- Normalfördelningens form påminner om en ringklocka (**bell-shaped distribution** på engelska).
- Kallas även för en Gaussisk fördelning (**Gaussian distribution** på engelska).

Normalfördelningen, forts.

- En standardiserad normalfördelning för variabeln $z = (y - \mu)/\sigma$:

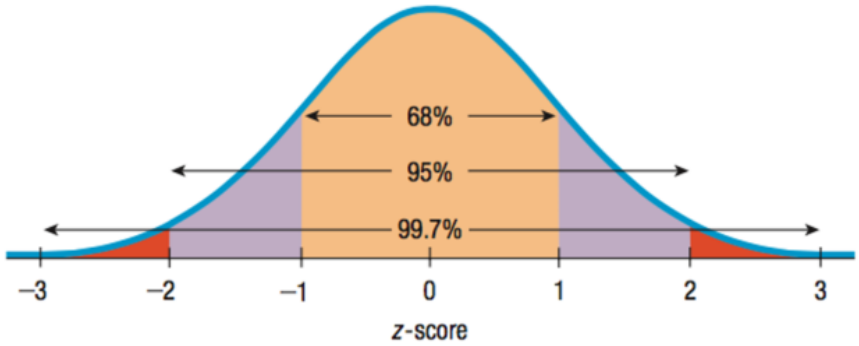


Figure 6: Figur 5.6 i De Veaux et al. (2021).

- En **teoretisk fördelning är en sannolikhetsmassa på 100% som är distribuerad över de olika utfall variabeln kan anta**.

- z faller inom (1.–3. kallas för **68–95–99.7 regeln**):
  1. intervallet $[-1, 1]$ (dvs ± 1 från mittpukten 0) med sannolikhet $\approx 68\%$.
  2. intervallet $[-2, 2]$ (dvs ± 2 från mittpukten 0) med sannolikhet $\approx 95\%$.
  3. intervallet $[-3, 3]$ (dvs ± 3 från mittpukten 0) med sannolikhet $\approx 99.7\%$.
  4. intervallet $(-\infty, \infty)$ med sannolikhet 100%.

Normalfördelningen, forts.

- Eftersom $z = (y - \mu)/\sigma \implies y = \mu + \sigma z$, har den teoretiska fördelningen för y medelvärde μ och standardavvikelse σ:

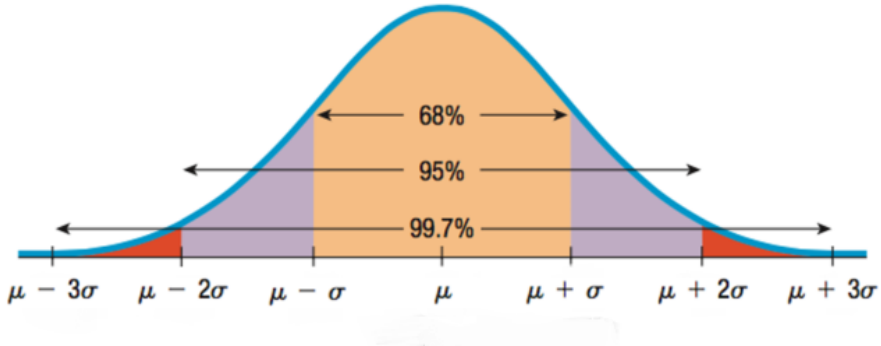


Figure 7: Figur 5.6 i De Veaux et al. (2021).

- y faller inom (1.–3. kallas för **68–95–99.7 regeln**):
  1. intervallet $[\mu - 1\sigma, \mu + 1\sigma]$ (dvs $\pm\sigma$ från μ) med sannolikhet $\approx 68\%$.
  2. intervallet $[\mu - 2\sigma, \mu + 2\sigma]$ (dvs $\pm 2\sigma$ från μ) med sannolikhet $\approx 95\%$.
  3. intervallet $[\mu - 3\sigma, \mu + 3\sigma]$ (dvs $\pm 3\sigma$ från μ) med sannolikhet $\approx 99.7\%$.
  4. intervallet $(-\infty, \infty)$ med sannolikhet 100%.

Population kontra stickprov

- Vi säger att y följer en normalfördelning med medelvärde μ och standardavvikelse σ. Betecknas $N(\mu, \sigma)$.
- Förr använde vi $\overline{y}$ för att beteckna **stickprovets medelvärde**, och s för att beteckna **stickprovets standardavvikelse**.
- Vi använder teoretiska fördelningar för att modellera **populationen**. En population har **parametrar** som beskriver populationen.
- μ är ett exempel på en **populationsparameter**. μ är populationens medelvärde. Även kallat väntevärde (mer om väntevärden i Del 2 av kursen).
- σ är ett annat exempel på en **populationsparameter**. σ är populationens standardavvikelse.
- $\overline{y}$ är stickprovets medelvärde.
- s är stickprovets medelvärde.

Population kontra stickprov, forts.

- Att **kunna urskilja en populationskvantitet (t.ex μ och σ) från en stickprovskvantitet (t.ex $\overline{y}$ och s)** är ett av de allra svåraste koncepten.
- Eftersom vi inte har tillgång till hela populationen, så kan vi kan **aldrig exakt mäta populationskvantiteter**, såsom μ eller σ.
- När vi tar ett stickprov, så kan vi **exakt beräkna stickprovskvantiteter**, såsom stickprovets medelvärde $\overline{y}$ eller stickprovets standardavvikelse s.
- **Inferentiell statistik** är ett ramverk för att utifrån ett stickprov från populationen **skatta/estimera populationsparametrarna**.
- Detta kommer att gås igenom i detalj i Del 2 av kursen.
- I Del 1 av kursen har vi alltid följande i åtanke:
  - Vi blandar aldrig ihop population och stickprov.
  - Speciellt blandar **vi aldrig ihop populationskvantiteter med stickprovskvantiteter**.

Normalfördelningen, forts.

- Tillbaka till normalfördelningen.
- Vi säger att y följer en normalfördelning med medelvärde μ och standardavvikelse σ. Betecknas $N(\mu, \sigma)$.
- Ett viktigt resultat ni kommer få lära er i del två av kursen är följande.
- Om y följer en $N(\mu, \sigma)$ så följer $z = \frac{y-\mu}{\sigma}$ en $N(0, 1)$.
- $N(0, 1)$ kallas för en **standard normalfördelning**.
- $z = \frac{y-\mu}{\sigma}$ kallas också z-värde. Gamla z-värdet standardiserade med hjälp av stickprovets medelvärde och standardavvikelse, dvs

$$z = \frac{y - \overline{y}}{s}.$$

- Det nya z-värdet standardiserar med hjälp av populationens medelvärde (väntevärde) och populationens standardavvikelse.

Normalfördelningen, forts.

- För att den teoretiska normalfördelningen ska vara en rimlig modell, bör den empiriska fördelningen data vara unimodal och någorlunda symmetrisk.
- Ett histogram är till stor hjälp för att avgöra detta.
- 68–95–99.7 regeln är också användbar för att se om normalfördelningen är en rimlig fördelning att anta.
- Vi räknar 68–95–99.7 intervallen på den empiriska fördelningen.
- Normalfördelningen rimlig om de empiriska intervallen ligger nära de motsvarande intervallen om data var normalfördelade, dvs $\overline{y} \pm x \cdot s$, $x = 1, 2, 3$.
- De empiriska intervallen kan beräknas genom `quantiles` funktionen i R som beräknar percentiler.
- Exempel: `quantiles(data, c(0.16, 0.84))` för 68% intervallet (16% svansar till vänster och höger)

Normalfördelningen, forts.

- Mätningar på handledsomkrets (i cm) från 250 män:

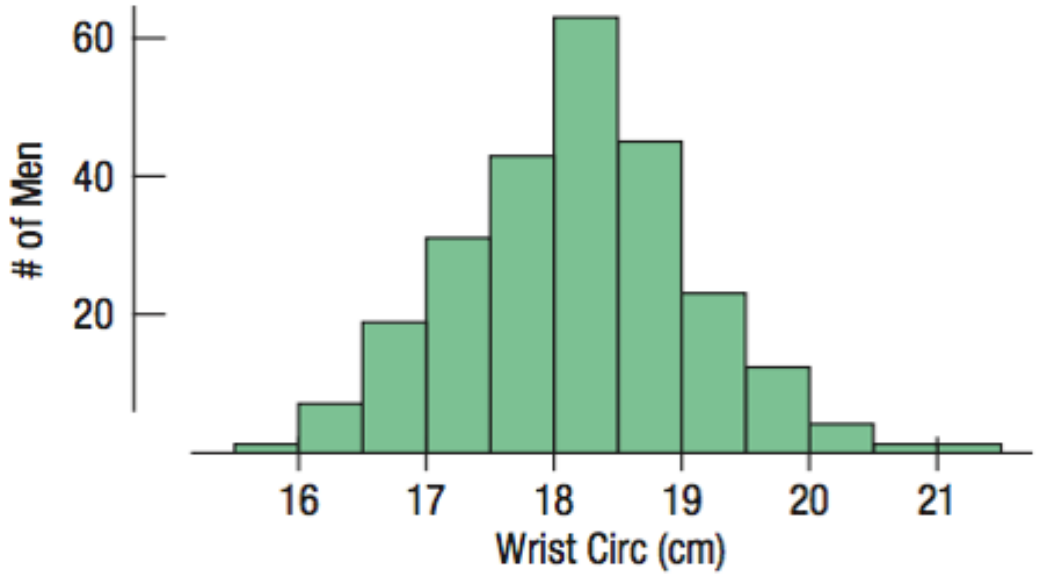


Figure 8: Figur från s.160 i De Veaux et al. (2021).

- Normalfördelningen är rimlig baserat på följande (beräknat med R):

	Percentil intervall	$\overline{y} \pm x \cdot s,\ x = 1, 2, 3$
68%	$[17.30, 19.10]$	$[17.31, 19.13]$
95%	$[16.50, 19.98]$	$[16.39, 20.05]$
99.7%	$[15.91, 21.21]$	$[15.48, 20.96]$

Normalfördelningen, forts.

- Tidigare frågade vi: Vad anses vara ett stort z värde? Är $z = 1$ stort? Eller $z = 2$?
- När z är normalfördelat vet vi svaret. Vi vet att z
  1. intervallet $[-1, 1]$ (dvs ± 1 från mittpukten 0) med sannolikhet $\approx 68\%$.
  2. intervallet $[-2, 2]$ (dvs ± 2 från mittpukten 0) med sannolikhet $\approx 95\%$.
  3. intervallet $[-3, 3]$ (dvs ± 3 från mittpukten 0) med sannolikhet $\approx 99.7\%$.
  4. intervallet $(-\infty, \infty)$ med sannolikhet 100%.
- $z = 1$ har 16% av sannolikhetsmassan till höger om sig. Hur mycket till vänster om sig?
- $z = 2$ har 2.5% av sannolikhetsmassan till höger om sig. Hur mycket till vänster om sig?
- $z = 1$ är ett ganska stort värde. $z = 2$ är ett mycket stort värde.

Normalfördelningen, forts.

- Betrakta normalfördelningen för variabeln $z = (y - \mu)/\sigma$ igen:

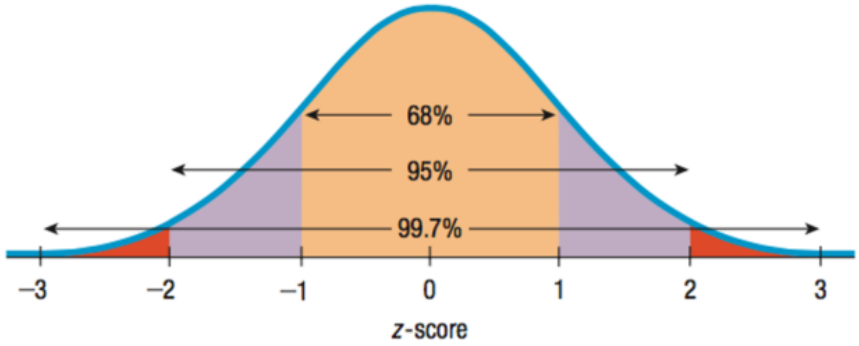


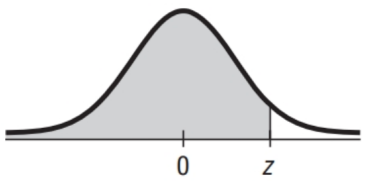
Figure 9: Figur 5.6 i De Veaux et al. (2021).

- Om z inte är -3, -2, -1, 0, 1, 2, eller 3, så kan vi inte utläsa exakta sannolikheter från figuren.
- Programvara används ofta. T.ex ger funktionen `pnorm` i R sannolikheter från normalfördelningen.
- `pnorm(1.2)` ger sannolikhetsmassan till vänster om $z = 1.2$ i en $N(0, 1)$ fördelning.
- Är `pnorm(1.2)` mindre eller större än 84%?
- Finns också normalfördelningstabeller man kan slå i.

Normalfördelningen, forts.

- Sannolikhetsmassorna till vänster om $z = 0.13$ och $z = 0.98$ när z följer en $N(0, 1)$?

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

Figure 10: Utdrag ur normalfördelningstabellen på s.1011 i De Veaux et al. (2021).

- Svar: 55.17% för $z = 0.13$ och 83.65% för $z = 0.98$.

Normalfördelningen, forts.

- Ibland är man intresserad av den omvända frågeställningen: Vilket z-värde ger att X% av sannolikhetsmassan ligger till vänster om z?
- Ett sådant z kallas för X-percentilen (**percentile** på engelska).
- Exempel: 90-percentilen är ett z-värde som har 90% av sannolikhetsmassan till vänster om sig (och 10% till höger om sig):

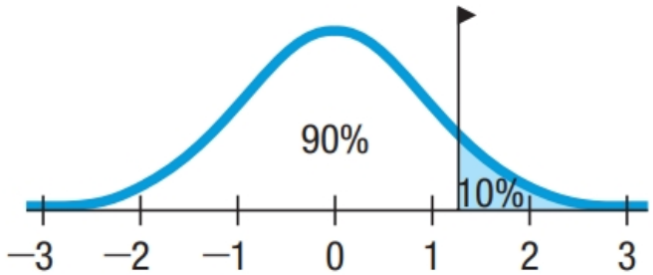
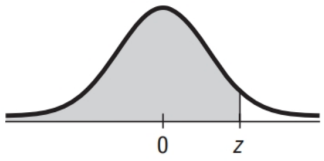


Figure 11: Figur från s.167 i De Veaux et al. (2021).

Normalfördelningen, forts.

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
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

Figure 12: Utdrag ur normalfördelningstabellen på s.1011 i De Veaux et al. (2021).

- $z \approx 1.28$ är den ungefärliga 90-percentilen i en standard normalfördelning.
- Funktionen `qnorm` i R ger percentiler från normalfördelningen.
- `qnorm(0.9)` ger ett exakt värde $z = 1.281552$ (upp till 6 decimaler).
- Villanis `tabeller` delas ut på tentan.

Normalfördelningen, forts.

- Vi har gått igenom 68–95–99.7 regeln och visat hur den kan användas för att stämma av om normalfördelningsantagandet är rimligt.
- Det finns andra verktyg man kan använda för att stämma av **normalfördelningens rimlighet** för ett specifikt datamaterial.
- Ett av de vanligaste är den så kallade normalfördelningsplotten (**normal probability plot** på engelska).
- "All models are wrong, but some are useful" – George E.P. Box.
- Kontentan är att vi aldrig kan förvänta oss att data följer en exakt normalfördelning. Vi får leva med små avvikelser (ibland mindre små...).

Normalfördelningen, forts.

- Exempel: Bensinförbrukning i mpg (miles per gallon) för en Nissan Maxima insamlat under 8 år av en av bokens författare.
- Normalfördelningsplot tillsammans med ett histogram:

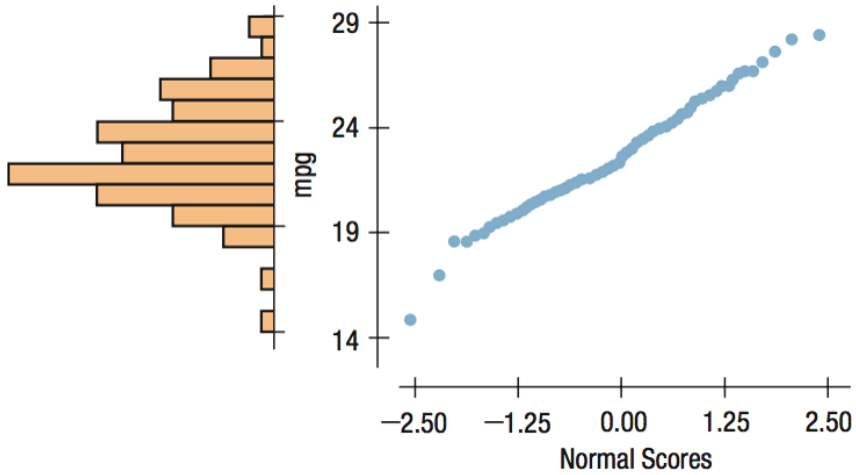


Figure 13: Figur 5.10 i De Veaux et al. (2021).

- Om observationerna **ligger längst en rät linje är de normalfördelade**.
- Två observationer i vänster svans är längre ut än vad som hade varit fallet om de hade varit normalfördelade. Normalfördelningen antas ändå rimlig.

Normalfördelningen, forts.

- Normalfördelningsplot tillsammans med ett histogram för exemplet med mäns vikter tidigare under föreläsningen:

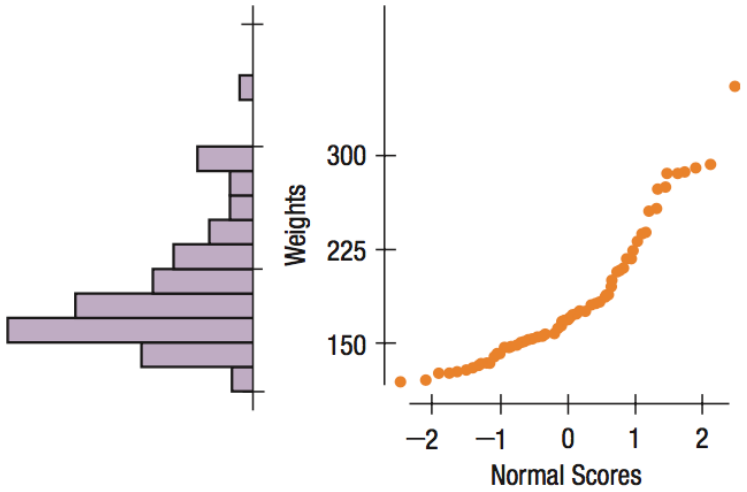


Figure 14: Figur 5.11 i De Veaux et al. (2021).

- Avviker stort från en rak linje. Normalfördelningsantagandet är helt orimligt.
- Inte ett oväntat resultat eftersom fördelningen är skev.
- Man kan testa att transformera data, till exempel genom en logaritmisk transform och studera normalfördelningsplotten för de transformerade data.

References I

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