Tentamen SDA1, del 1, lösningsförslag

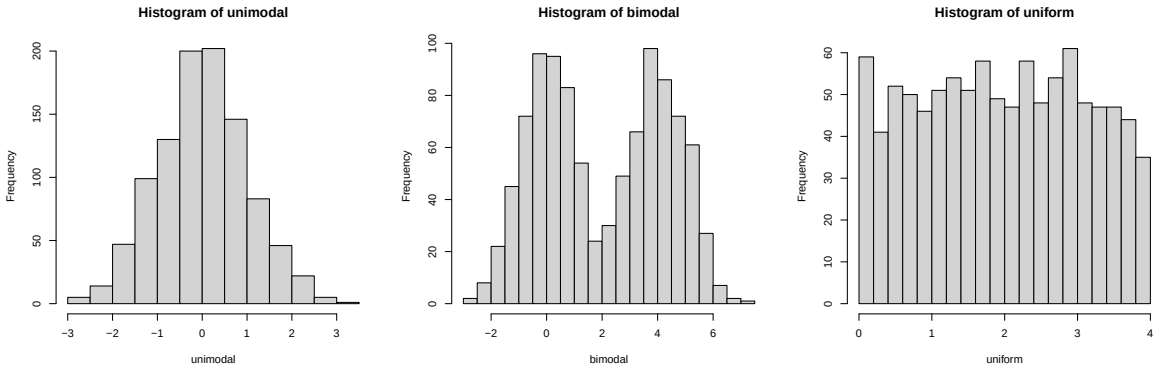
Uppgift 1

(a.) Nämn två typer av diagram som kan användas för att illustrera fördelningen för kategoriska variabler. Nämn även två typer av diagram som kan användas för att illustrera fördelningen för numeriska variabler. (5p)

Fördelningen för en kategorisk variabel kan illustreras exempelvis i ett stapeldiagram eller ett pajdiagram.

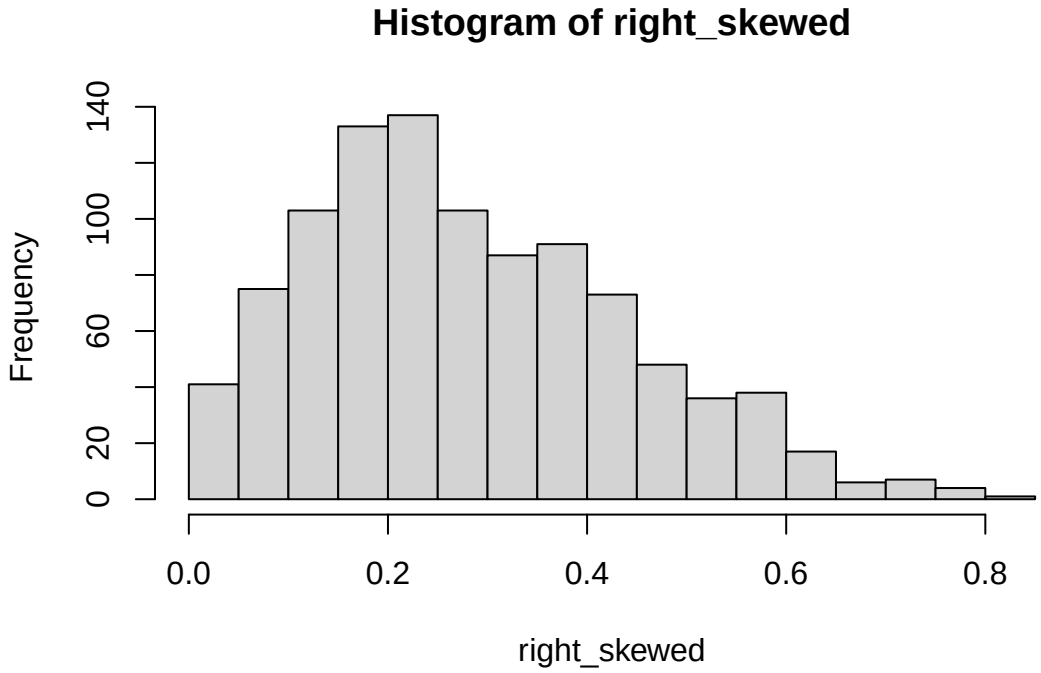
Fördelningen för en numeriska variabel kan illustreras exempelvis i ett histogram eller i ett låddiagram.

(b.) Rita en unimodal, en bimodal, och en uniform fördelning. Skriv ut vilken graf som illustrerar vilken fördelning. Det behöver inte vara snyggt ritat, men för poäng måste det synas tydligt vilken typ av fördelning varje graf illustrerar. (5p)



(c.) Tre centralmått för en fördelning är typvärde (mode), median och medelvärde (mean). Om en fördelning är skev åt höger (right skewed), vilket av dessa centralmått kan förväntas ha högst värde? Motivera ditt svar, och rita gärna en bild. (5p)

Illustration av en fördelning som är skev åt höger:



Vi kan förvänta oss att medelvärdet är högst. Att fördelningen är skev åt höger betyder att det finns fler värden som är mycket höga än vad det finns värden som är mycket låga. Enstaka värden som är mycket höga driver upp genomsnittet, men de driver inte upp medianen eller typvärdet.

(d.) Nämn en fördel med studier i form av experiment jämfört med observationsstudier. (5p)

Om en studie är upplagd som ett experiment är det möjligt att dra slutsatser om kausala samband. Det beror på att vi i ett experiment kan bestämma hur förutsättningarna ska se ut.

Uppgift 2

(a.) När vi anpassar en enkel linjär regressionsmodell, det vill säga när vi bestämmer vilka värden b_0 och b_1 ska ha, gör vi det med hjälp av Minsta Kvadratmetoden (Least Squares Method). Vad menas med att vi använder Minsta Kvadratmetoden? (5p)

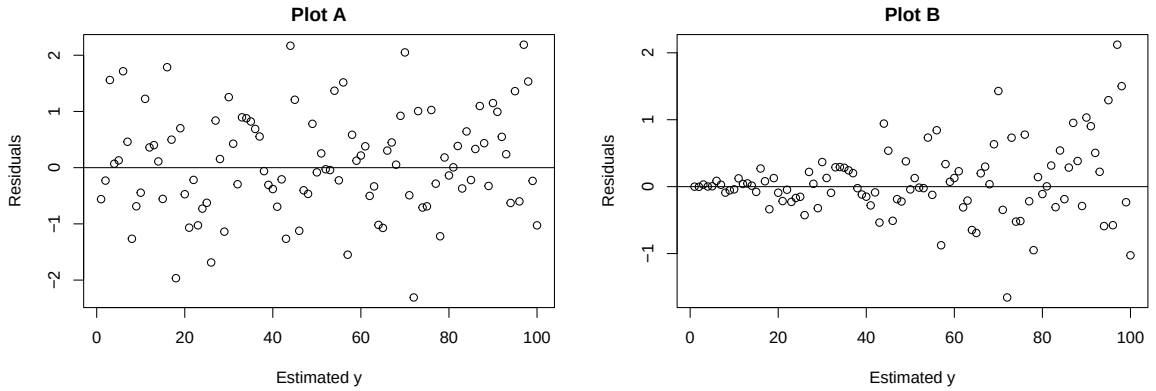
Vi sätter de värden på b_0 och b_1 som ger den minsta summan av de kvadrerade residualerna, dvs vi väljer den regressionslinje som minimerar

$$\sum_{i=1}^{n} e^2$$

(b.) Vad menar vi när vi säger att sambandet mellan responsvariabeln och förklaringsvariabeln i en enkel linjär regressionsmodell inte är statistiskt signifikant? (5p)

Vi menar att det samband som vi ser i vår data mycket väl kan vara ett resultat av slumpen, och att det därmed inte är ett samband som gäller generellt för den population som vi är intresserade av.

(c.) Nedan ser vi residualgrafer för två olika linjära regressionsmodeller. Vilken av graferna kommer från en modell där vi ser att alla våra modellantaganden *inte* är uppfyllda? Vilket modellantagande är inte uppfyllt? (5p)



Graf **B** visar en residualplot från en funktion som inte uppfyller antagandet om att residualerna ska ha konstans varians. Vi ser att variansen ökar med $\hat{y}$ (Estimated y).

(d.) En regressionsmodell bör varken ha för hög eller för låg flexibilitet. Förklara varför. (5p)

Med för låg flexibilitet klarar inte modellen att fånga datans övergripande mönster. Med för hög flexibilitet anpassar sig datan allt för mycket till de enskilda observationerna vilket gör den sämre lämpad att göra prediktioner på ny data.

Uppgift 3

(a.) Vad kallas denna typ av tabell, och vilken typ av variabler är det vi studerar med den här sortens tabell?(4p)

Det är en korstabell, eller en simultanfördelningstabell. Den används för att studera kategoriska variabler.

(b.) Räkna ut marginalfördelningarna (marginal distributions) för båda variablerna. Uttryck marginalfördelningarna i procent. Tolka marginalfördelningarna. (4p)

Vi börjar med att räkna ut det totala antalet observationer (dvs antalet deltagare i studien):

$178 + 121 + 70 + 57 = 426$

Därefter räknar vi ut marginalfördelningen av variabeln Grupp, uttryckt i antal:

Läkemedel: $178 + 70 = 248$

Placebo: $121 + 57 = 178$

Uttryckt i procent blir det:

Läkemedel: $248/426 = 0.58 = 58\%$

Placebo: $178/426 = 0.42 = 42\%$

Vi räknar ut marginalfördelningen av variabeln Effekt, uttryckt i antal:

Ja: $178 + 121 = 299$

Nej: $70 + 57 = 127$

Uttryckt i procent blir det:

Ja: $299/426 = 0.70 = 70\%$

Nej: $127/426 = 0.30 = 30\%$

Tolkning: 58% av deltagarna fick läkemedlet och 42% fick placebobehandling. 70% av deltagarna mådde bätte efter behandlingen, och 30% mådde inte bättre.

(c.) Räkna ut fördelningen av variabeln *Effekt* betingad (conditioned) på variabeln *Grupp*. Ange de betingade fördelningarna i procent. (8p)

```
m <- matrix(c(178, 121, 70,57), ncol=2)
m2 <- m / rowSums(m)
colnames(m2) <- c("Ja", "Nej")
rownames(m2) <- c("Läkemedel", "Placebo")
m2
```

```
                Ja       Nej
Läkemedel 0.7177419 0.2822581
Placebo   0.6797753 0.3202247
```

(d.) Tolka resultatet från deluppgift c. Vad säger resultatet om läkemedlets effekt? (4p)

Ungefär 72% av dem som fick läkemedlet mådde bättre efter behandlingen. Av dem som fick placebo var det ungefär 68% som mådde bättre.

Skillnaden är inte särskilt stor.

Uppgift 4

En butiksägare väger de äpplen som kommit med den senaste leveransen. Det visar sig att äpplenas vikt följer en normalfördelning med medelvärdet 200 gram och standardavvikelsen 40 gram.

(a.) Leverantören hävdar att nästan alla äpplen i leveransen väger mer än 140 gram. För att undersöka hur väl detta stämmer, använd normalfördelningen och räkna ut hur stor andel av alla äpplen i leveransen som väger mer än 140 gram. (5p)

```
average <- 200
s <- 40
z <- (140 - 200) / 40
z
```

```
[1] -1.5
```

```
pnorm(q=z)
```

```
[1] 0.0668072
```

Ungefär 6.7 procent av alla äpplen vägen mindre än 140 gram, och 93.3% väger mer.

(b.) Handlaren sorterar ut de 10 procent äpplen som väger mest, för att sälja dessa dyrare. Hur mycket måste ett äpple väga för att tillhöra dessa 10 procent, dvs vad måste ett äpple väga som minst för att vikten ska ligga över den 90:e percentilen? (5p)

```r
z <- qnorm(p=0.9)
z
```

```
[1] 1.281552
```

```r
average + z * 40
```

```
[1] 251.2621
```

Ett äpple måste väga drygt 251 gram för att tillhöra de 10 procent tyngsta.

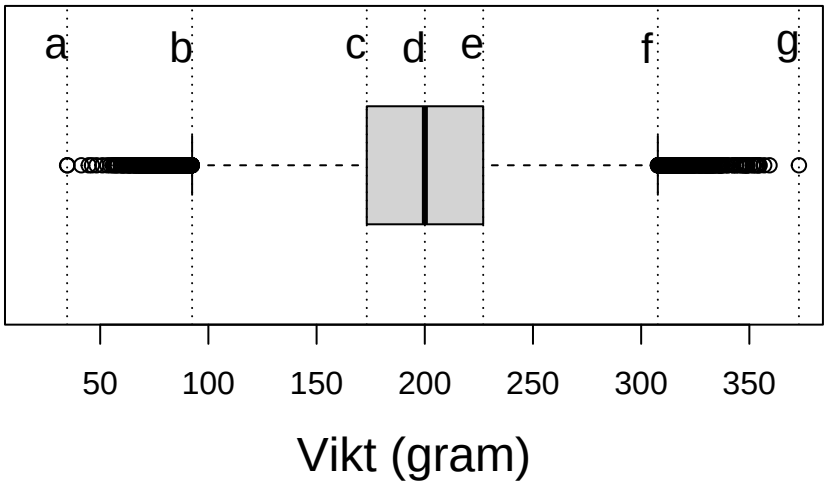
(c.) Hur stor andel av alla äpplen har en vikt som avviker med högst 40 gram från medelvärdet? Hur stor andel har en vikt som avviker med högst 80 gram från medelvärdet? (5p)

En standardavvikelse är 40 gram, och vi vet att 68% av alla observationern avviker från medelvärdet med högst en standardavvikelse.

Två standardavvikelser är 80 gram, och vi vet att 95% av alla observationer avviker med medelvärdet med högst 2 standardavvikelser.

Uppgiften kan också lösas med hjälp av normalfördelningstabellen enligt samma princip som deluppgift **a**.

(d.) Låddiagrammet (box plot) är ett sätt att illustrera viktfördelningen för de äpplen som levererades. Vad betecknar var och en av bokstäverna a, b, c, d, e, f och g i diagrammet? (Du behöver inte läsa av siffrorna eller räkna ut något i denna deluppgift) (5p)



a: Den minsta observationen. b: Gränsen för outliers med låga värden. c: 1:a kvartilen. d: Medianen. e: 3:e kvartilen. f: Gränsen för höga outliers. g. Den största observationen.

Uppgift 5

En badvakt har samlat in data över antalet personer som badar på en badstrand dag för dag. Han har använt denna data till att anpassa regressionsmodellen

$$\hat{y} = 29 + 3.2x,$$

där x är vattentemperaturen i grader Celsius en viss dag och $\hat{y}$ är det estimerade antalet personer som badar.

(a.) Tolka modellens koefficienter (coefficients), dvs interceptet och lutningskoefficienten (slope coefficient). (4p)

Om vattentemperaturen är 0 grader estimerar modellen att 29 personer badar den dagen. (Tolka dock interceptet med försiktighet och inte bokstavligt som en prediktion.) För varje ytterligare grad Celsius estimerar modellen ytterligare 3.2 badgäster.

(b.) En viss dag estimerade badvaktens modell att 95 personer skulle bada på stranden. Ungefär vilken temperatur var det den dagen? (4p)

Om $\hat{y} = 95$ får vi ekvationen

$$95 = 29 + 3.2x$$

Om vi löser ekvationen får vi $x = (95 - 29)/3.2 = 20.625$, så temperaturen den dagen var ungefär 20.6 grader.

(c.) Vi har följande information:

$r_{xy} = 0.4, s_x^2 = 37.21, s_y^2 = 2401, \bar{x} = 19, \bar{y} = 90.$

Förklara vad r_{xy}, s_x^2 och s_y^2 står för.

Verifiera sedan med hjälp av den givna information att $b_0 \approx 29$ och att $b_1 \approx 3.2$. (4p)

r_{xy} är korrelationen mellan vattentemperaturen och antalet badande en dag. s_x^2 är variansen för temperaturen. s_y^2 är variansen för antalet badande.

$s_x = \sqrt{s_x^2} = \sqrt{37.21} = 6.1$ $s_y = \sqrt{s_y^2} = \sqrt{2401} = 49$

$b_1 = r_{xy} \cdot \frac{s_y}{s_x} = 0.4 \cdot \frac{49}{6.1} = 3.2131 \approx 3.2$ $b_0 = \bar{y} - b_1 \cdot \bar{x} = 90 - 3.2131 \cdot 19 = 28.9511 \approx 29$

(d.) Vi lägger till dummy-variabeln *regn*, som har värdet 1 om det regnade en viss dag och annars 0. Modellen blir nu.

$$\hat{y} = 72 + 1.2x_1 - 26x_2,$$

där x_1 är vattentemperaturen i grader Celsius och x_2 är dummyvariabeln för regn.

Tolka modellens samtliga tre koefficienter b_0, b_1 och b_2. (4p)

b_0: Modellen estimerar 72 personer som badar om temperaturen är 0 grader och det inte regnar. b_1: Modellen estimerar att ytterligare 1.2 personer badar för varje ytterligare grad, **givet ett visst värde på variabeln regn**. $b_{2\_}$ Modellen estimerar att 26 personer färre badar en dag med regn, **givet en viss vattentemperatur**.

För att svaret ska vara korrekt måste det specificeras att tolkningen av koefficienten för en variabel förutsätter att vi betraktar värdet på den andra variabeln som givet, dvs tolkningen gäller när den andra variabeln hålls konstant.

(e.)

Vi har följande information om en regressionsmodell:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 30000$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 9000$$

Hur stor del av responvariabelns variation förklarar regressionsmodellen? (4p)

R-squared mäter hur stor del av respondvariabelns variation som förklaras av regressionsmodellen. Vi ska alltså räkna ut R-squared.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{9000}{30000} = 0.7$$

Det betyder att regressionsmodellen förklarar 70% av responsvariabelns variation.