Lösningsförslag till tenta 20240927

Uppgift 1

a

Deskriptiv statistik handlar om att beskriva ett datamaterial, exempelvis med hjälp av diagram eller tabeller. Deskriptiv statistik kan till exemepel illustrera samband mellan variabler, eller visa hur värden på en enskild variabel fördelar sig i datamaterialet.

Inferens handlar om att dra generella slutsatser utifrån av datamaterialet. Exempel: Baserat på en enkät som ett antal invånare svarar på, vilka slutsatser drar vi om vad alla invånare tycker? Vi kan uttrycka det som att vi drar slutsatser om en population med hjälp av ett stickprov (ett sample).

b

Vi har följande frekvenstabell:

```
          kategori antal
1        Telefoner    56
2          Datorer    32
3     TV-apparater    12
4 Hushållsmaskiner     7
5           Övrigt     4
```

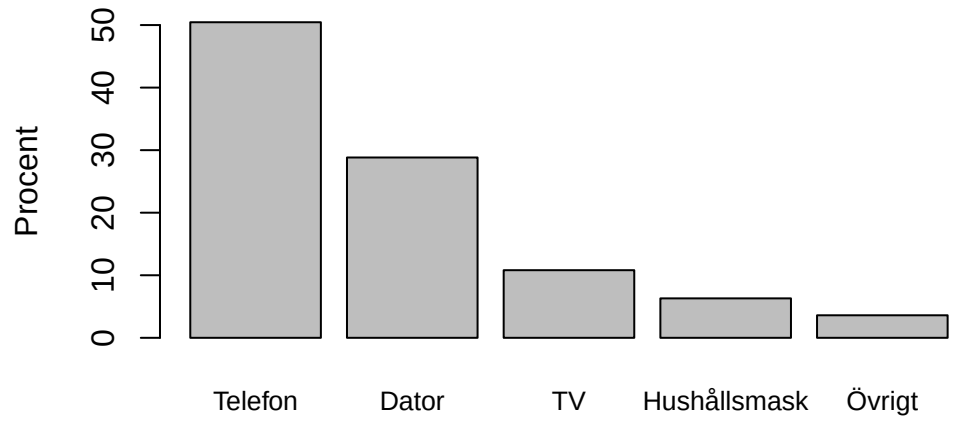
För att omvandla frekvenstabellen till en relativ frekvenstabell räknar vi först ut det totala antalet observationer. Sedan delar vi varje antal i frekvenstabellen med det totala antalet.

```
totalt <- 56 + 32 + 12 + 7 + 4
totalt
```

```
[1] 111
```

```
          kategori relativ_frekvens_procent
1        Telefoner                50.450450
2          Datorer                28.828829
3     TV-apparater                10.810811
4 Hushållsmaskiner                 6.306306
5           Övrigt                 3.603604
```

Eftersom *kategori* är en kategorisk variabel kan vi använda ett stapeldiagram eller ett pajdiagram. Här använder vi ett stapeldiagram.



c

Vi får medelvärdet genom att summera alla värden och dela med antalet observationer. Om antal poäng är x blir medelvärdet

$$\bar{x} = \frac{56 + 72 + 82 + 17 + 43 + 43 + 51}{7} = 52$$

Typvärdet (mode) är det värde som förekommer flest antal gånger. I den här fördelningen är 43 det värde som förekommer oftast, då det förekommer 2 gånger.

Medianen är det mittersta värdet om vi sorterar alla värden i storleksordning. Våra sorterade värden är

$17, 43, 43, 51, 56, 72, 82$

Det mittersta värdet är 51, vilket alltså är medianen.

d

Studentens påstående är *inte* korrekt. Den tredje kvartilen är ett värde som är större än ungefär 75 procent av observationerna och mindre än ungefär 25 procent av observationerna. En student med 43 poäng har en poäng som är lägre än medianen, med andra ord lägre än den andra kvartilen, så poängen ligger långt under den tredje kvartilen.

Uppgift 2

a

En residual är skillnaden mellan ett observerat värde och det skattade värdet för samma observation.

$e = y - \hat{y}$, där e är residualen, y det observerade värdet och $\hat{y}$ det skattade värdet.

Minsta kvadraten-metoden säger att vi väljer den regressionsmodell, dvs de värden på våra regressionskoefficienter, som ger minsta möjliga summa av de kvadrerade residualerna. Vi minimerar alltså

$\sum_i e_i^2$.

b

Om vi utvärderar modellen på samma data som vi använder för att anpassa modellen finns det en risk att en modell som ser bra ut i själva verket är överanpassad. Det innebär att modellen har anpassat sig väl till vår träningsdata, men presterar dåligt när vi använder den för att göra estimationer med *ny* data. Modellen är då inte generaliserbar.

Om vi istället anpassar modellen med träningsdata och utvärderar den med separat testdata har vi använt olika data för anpassning respektive utvärdering. Då kan vi förvänta oss ett resultat av utvärderingen som påminner mer om det vi får när vi gör prediktioner med ny data.

c

Om sambandet inte är linjärt kan det ändå vara möjligt att använda en enkel linjär regressionsmodell efter att ha transformerat en eller båda variablerna. Vi kan använda Tukey's cirkel för att hitta en lämplig transformation givet det samband som vi ser mellan variablerna. Om vi lyckas få ett linjärt samband mellan de transformerade variablerna kan vi anpassa en enkel linjär regressionsmodell till dem.

d

Residualgrafen säger att vi inte har konstant varians, så det modellantagandet är inte uppfyllt. Normalfördelningsgrafen följer på ett ungefär en rät linje, så antagandet om att residualerna är normalfördelade kan vi betrakta som uppfyllt. (Svaret att residualerna inte är tillräckligt normalfördelade är också ok, om det framgår av svaret att normalfördelade residualer ger en rät linje i normalfördelningsgrafen.)

Uppgift 3

```r
m <- matrix(c(31500, 33300, 28800, 2533, 3362, 2277), ncol=2)
colnames(m) <- c("Ja", "Nej")
rownames(m) <- paste("Högskola", 1:3)
m
```

```
             Ja  Nej
Högskola 1 31500 2533
Högskola 2 33300 3362
Högskola 3 28800 2277
```

a) Hur många studenter svarade på enkätfrågan?

```r
sum(m)
```

```
[1] 101772
```

101 772 studenter svarade på enkäten.

b) Räkna ut marginalfördelningarna i procent för båda variablerna.

```r
rowSums(m) / sum(m)
```

```
Högskola 1 Högskola 2 Högskola 3
 0.3344044  0.3602366  0.3053590
```

```r
colSums(m) / sum(m)
```

```
        Ja        Nej
0.91970287 0.08029713
```

Tolkning:

Avrundat till heltal hade 33% av studenterna i undersökningen examen från högskola 1, 36% från högskola 2 och 31% från högskola 3.

Av alla studenter i undersökningen hade ungefär 92% en anställning ett år efter examen och ungefär 8% inte en anställning.

c)

Vi vill ha fördelningen av variabeln *Anställning* betingad på variabeln *Högskola*. Då kan vi se hur väl var och en av högskolorna har lyckats, och jämföra resultaten.

d) Räkna ut den betingade fördelning som du föreslog i deluppgift c

```
m / rowSums(m)
```

```
                  Ja        Nej
Högskola 1 0.9255722 0.07442776
Högskola 2 0.9082974 0.09170258
Högskola 3 0.9267304 0.07326962
```

e) Tolka resultatet från deluppgift d.

Det är ingen större skillnad mellan skolorna. Av examinerade från högskola 2 har drygt 90 procent fått en anställning ett år efter examen. Av examinerade från högskola 1 och högskola 3 har lite mer än 92 procent fått en anställning, så de har lyckats en aning bättre än högskola 2.

Uppgift 4

a) Tolka modellens koefficienter

b_1: För varje miljon en kund spenderar under ett år får kunden ytterligare 0.2 procentenheter i rabatt.

b_0: En kund som spenderar noll kronor under ett år får en rabatt på 0.05 procent. Det ska inte tolkas bokstavligt.

b) Förklara notationen och verifiera regressionskoefficienterna

Vi har följande information: $r_{xy} = 0.25, s_x = 2, s_y = 1.6, \bar{x} = 4.8, \bar{y} = 1.01$

r_{xy} är korrelationskoefficienten för variablerna x och y. s_x och s_y är standardavvikelsen för x respektive y. $\bar{x}$ och $\bar{y}$ är variablernas medelvärden.

För att verifiera b_1 kan vi använda formeln

$$b_1 = r_{xy}\frac{s_y}{s_x} = 0.25 \cdot \frac{1.6}{2} = 0.2$$

För att verifiera b_0 kan vi använda $b_0 = \bar{y} - b_1\bar{x} = 1.01 - 0.2 \cdot 4.8 = 0.05$

c) Vad blir rabatten för en kund som handlar för 240 miljoner kronor?

Vi sätter in 240 som värdet på x, och då får vi

$\hat{y} = 0.05 + 0.2 \cdot 240 = 48.05.$

Kunden skulle enligt modellen få ungefär 48 procents rabatt. Det verkar orimligt mycket, och eftersom köp för 240 miljoner kronor ligger långt från de värden som vi har i vår data (1.2 till 42 miljoner) kan vi inte utgå från att modellen är användbar i det här fallet.

d) Tolka regressionskoefficienterna i den multipla regressionen

b_0: En kund som spenderar noll kronor och som inte är ny får -0.01 procent i rabatt. Kan inte tolkas bokstavligt.

b_1: En kund får ytterligare 0.15 procentenheter i rabatt per spenderad miljon, *givet* värdet på x_2. Om vi exempelvis har två kunder som båda är nya förväntar vi oss att den kund som spenderar en miljon mer har 0.15 procentenheter större rabatt.

b_2: Givet att en kund spenderar en viss summa så får nya kunder ytterligare 0.09 procentenheter i rabatt. Om vi har två kunder, en ny kund och en gammal, som båda spenderar lika mycket, då förväntar vi oss att den nya kunden har 0.09 procentenheter högre rabatt.

e) Råkna ut R-kvadrat

Vi ser att SST är 22300, och att SSE minus SSE är 19251. Det betyder att SSR är 19251, och vi kan använda formeln

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{19251}{22300} \approx 0.86$$

R-kvadrat är ungefär 0.86.

Uppgift 5

a) Hur stor andel hade en lägre kostnad än 700 kronor?

Vi vet att $\bar{x} = 570, s_x = 120$. Vi vet också att x följer en normalfördelning.

700 kronor är en kostnad som motsvarar följande z-värde:

$$z = \frac{x - \bar{x}}{s_x} = \frac{700 - 570}{120} \approx 1.08$$

I normalfördelningstabellen kan vi se att 85.99% av alla observationer har ett z-värde som är lägre än 1.08. Ungefär 86% av alla kunder hade allså en lägre kostnad än 700 kronor.

b) Vilken elkostnad motsvarar den 10:e percentilen?

Vi kan använda normalfördelningstabellen för att hitta det z-värde som morsvarar den 10:e percentilen, dvs z-värdet för en observation som är större än 10% av alla observationer.

Den 10:e percentilen finns inte i tabellen, så vi tittar i stället på den 90:e percentilen. Det värde som ligger närmast 0.9 i tabellen är 0.8997, som motsvarar ett z-värde på 1.28. Det betyder att z-värdet som motsvarar den 10:e percentilen är ungefär -1.28.

Därefter räknar vi ut vilken kostnad (vilket värde på x) som motsvarar z-värdet -1.28.

$x = \bar{x} + z \cdot s_x = 570 - 1.28 \cdot 120 \approx 416$

En kostnad som ligger vid den 10:e percentilen är alltså ungefär 416 kronor.

c) Räkna ut kvartilavståndet med hjälp av normalfördelningen

Vi börjar med att använda normalfördelningstabellen för att hitta de z-värden som motsvarar Q3 (75:e percentilen) och Q1 (25:e percentilen).

För Q3 är z = 0.67 (att använda 0.68 är också ok), och för Q1 blir då z-värdet -0.67.

Det betyder att

$Q3 = 570 + 0.67 \cdot 120 \approx 650$

$Q1 = 570 - 0.67 \cdot 120 \approx 490$

$IQR = Q3 - Q1 = 650 - 490 = 160$

Kvartilavståndet blir alltså ungefär 160. Med z-värdet 0.68, eller med andra avrundningar, kan svaret skilja sig en aning.

d) Räkna ut kvartilavståndet med hjälp av låddiagrammet.

log(Q3) motsvarar d i diagrammet, och log(Q1) motsvarar b.

Det betyder att

$\log(Q3) = 6.48 \implies Q3 = e^{6.48} \approx 652$

$\log(Q1) = 6.2 \implies Q1 = e^{6.2} \approx 493$

$IQR = Q3 - Q1 = 652 - 493 = 159$

Vi får alltså kvartilavståndet till 159, vilket är ungefär samma resultat som i deluppgift (c). Resultatet kan skilja sig något beroende på hur vi avrundar talen i beräkningarna.