Petter Mostad
Matematisk Statistik
Chalmers och GU

MSG830 Statistisk analys och experimentplanering

Tentamen 17 Januari 2014, 8:30 - 12:30
Examinator: Petter Mostad, telefon 0707163235,
kommer till tentamenslokalen 9:30 och 11:30
Tillåtna hjälpmedel: Kalkylator
Antal poäng totalt: 30. För godkänd resultat krävs minst 12 poäng

1. Figur 1 visar ett scatterplot av bivariata data, där variablerna x och y har observerats för ett antal objekter. Korrelationen mellan x och y är ett av talen $-1.18, -0.87, 0.66, 44.0$. Argumentera ut i från figuren vilket av talen som anger korrelationen. (1 poäng).

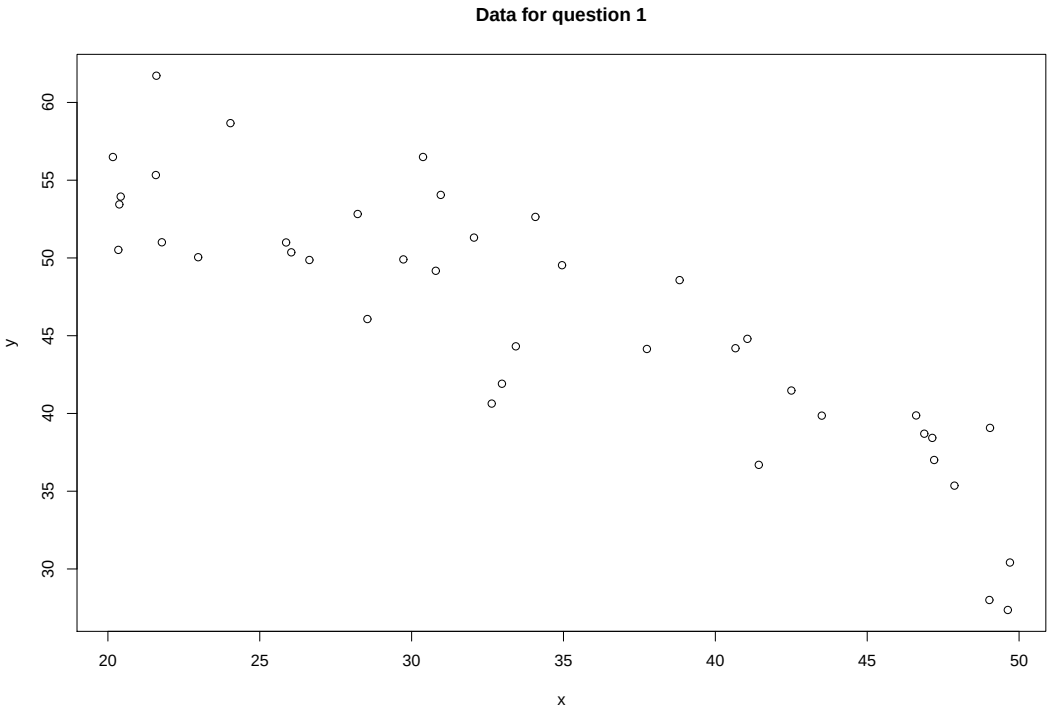

Figure 1: Data för fråga 1

2. En sjukdom X förekommer hos 8 personer varje år, i medelvärde, i en stad med en halv miljon personer. Vi antar sjukdomen förekommer i varje person oberoende av andra personer.

   (a) Vad är sannolikheten att exakt 5 personer får sjukdomen under ett år? (1 poäng)

(b) Vad är (ungefärlig) sannolikhet att 15 eller flera personer får sjukdomen under ett år? (2 poäng)

(c) Johan mår inte bra och är orolig för att han fått X. Han har symptom S, och har hittat på intenet at 75% av alla pasienter med sjukdom X har symptom S, medan bara 0.1% av personer som inte har X har symptom S. Vad är sannolikheten för att Johan har sjukdom X? (2 poäng)

(d) Johan får höra att sjukdomen har observerats hos 18 personer i hans stad detta året, inkluderat hos en av hans närmsta vänner. Diskutera om detta skulle kunna influera sannolikheten att Johan har X. (1 poäng)

3. Aila har gjord 40 observationer av produktionsmängden från en process under normala förhållanden. Observationerna värkar vara normalfördelade och har ett medelvärde på 3.39 och en stickprovsvarians på 0.2143. Hon har också gjord 40 observationer av produktionsmängden från samma process men under experimentella förhållanden. Dessa observationer värkar också vara normalfördelade och har ett medelvärde på 3.62 och en stickprovsvarians på 0.4501.

   (a) Vi antar att data kommer från normalfördelningar med samma fördelningsvarians. Beräkna ett 95% kredibilitetsintervall för skillnaden mellan väntevärden för fördelningarna. (2 poäng)

   (b) Baserat på beräkningarna i (a), är det rimligt att anta att de två normalfördelningar som data kommer från har samma väntevärde, så att de är identiska? (1 poäng)

   (c) Är antagandet i (a) om att fördelningsvarianserna är de samma rimligt? Använd en hypotestest för att svara på frågan. (1 poäng)

4. Om $x_1, x_2, \ldots, x_n$ är ett stickprov från en fördelning och $y_1, y_2, \ldots, y_m$ är ett stickprov från en annan fördelning, så kan man använda en Mann-Whitney U hypotestest för att testa om dessa två fördelningarna är den samma fördelningen. Berätta om idén bakom denna testen. Förklara också gärna hur testen genomförs, till exempel hur teststatistikan beräknas. (2 poäng)

5. Carlos studerar om växten av hans potatisodling beror på vilken pesticid han använder, eller vilken gödsel han använder. Han har valet mellan pesticiderna A och B, och mellan gödseltyperna X, Y, Z, och W. För varje kombination väljer han på ett randomiserat sätt 3 små odlingsområden för att testa ut kombinationen. Detta betyder att han har totalt 24 observationer av mäng producerat potatis. Den första tabellen under visar observationerna, medan den andra visar olika medelvärden: Varje cell visar medelvärdet för observationerna med denna kombinationen av faktorer, medan medelvärden för varje rad och kolumn också finns med, samt det totala medelvärdet för alla observationer. Stickprovsvariansen för alla data är 29.36775.

	A	B
X	41.6	40.3
	38.7	37.2
	37.1	43.1
Y	41.5	48.4
	33.9	47.2
	44.7	51.2
Z	49.2	40.5
	38.3	52.2
	43.3	46.1
W	41.9	42.6
	29.1	41.0
	38.1	38.2

	A	B	Medelvärde
X	37.55833	41.775	39.66667
Y	42.375	46.59167	44.48333
Z	42.825	47.04167	44.93333
W	36.375	40.59167	38.48333
Medelvärde	39.78333	44	41.89167

Gör en ANOVA tabell för Carlos. Beskriv konklutionerna som kan göras på grundlag av tabellen. (3 poäng).

6. Alexey planerar ett försök där han vill studera skillnaderna i försurning av sjöar i Sverige och Finland. Under våren mätar han surhetsgraden i 5 sjöar i Göteborgsområden. På sommaren åker han på semester till finska Lappland, där han hinner med att mäta surhetsgraden i 5 sjöar i området. Beskriv åtminstone två allvarliga brister i Alexey's försöksplan, och förklara varför detta är brister. (2 poäng)

7. Skriv ned en fraktionell faktoriell försöksplan ("fractional factorial design") med 4 försök och 3 faktorer, i form av en tabell med "+" och "-" (1 poäng).

8. Ge definitionen av begräppen "typ 1 fel" och "typ 2 fel" inom hypotestesting (1 poäng).

9. Jonathan studerar olika bi-artar, speciellt arterna X, Y, och Z. I vanliga biotoper är 34% av bin av arten X, 22% av arten Y, och 15% av arten Z, medan resterande bin är av andra arter. Jonathan studerar nu en speciell biotop, där han har artsbestämd 40 bin. Av dessa är 14 av arten X, 12 av arten Y, 3 av arten Z, och resterande av andra arter. Gör ett hypotestest och beräkna ett p-värde, för att ta reda på om den speciella biotopen skiljer sig från vanliga biotoper när det gäller förekomster av bi-arter. (2 poäng).

10. Anna studerar en speciell växt som blomstrar i kyla, och speciellt vill hon studera sannolikheten p för att växten blomstrar efter 10 dagar med kylbehandling. Preliminära studier

har visat att p är någon stans nära 0.3 - 0.5. Mera precist använder Anna aprioriförde lningen $p \sim \text{Beta}(4, 6)$.

(a) Anna kylbehandlar 72 växter, och av dessa blomstrar 27 efter 10 dagar. Beräkna posteriorifördelningen för p givet dessa data. (1 poäng)

(b) Beräkna ett ungefärligt 95% kredibilitetsinterval för p givet dessa data. (1 poäng)

(c) Om $X \sim \text{Beta}(\alpha, \beta)$, så är väntevärdet för X $\frac{\alpha}{\alpha+\beta}$ och fördelningsvariansen $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Beräkna väntevärdet och fördelningsvariansen för posteriorifördelningen för p. (1 poäng)

(d) Precis som Binomialfördelningen och Poissonfördelningen kan approximeras med Normalfördelningen, så kan även Betafördelningen approximeras på detta sätt. Använd en normalapproximation för att beräkna ett 95% kredibilitetsinterval för posteriorifördelningen för p. (2 poäng)

11. I en studie blev 5 variabler (A, B, C, D, och E) observerat för 50 städer. Principalkomponentanalys (PCA) blev använd t för att visualisera och studera data med användning av R-kommandon under:

```
> result <- prcomp(citydata)
> result
Standard deviations:
[1] 25.752948 11.757476  2.702226  2.049107  1.730706

Rotation:
           PC1         PC2         PC3         PC4         PC5
A -0.017007826  0.02704027 -0.27503336 -0.68832445 -0.67048167
B -0.876886331  0.47910893 -0.01846684  0.02656714  0.02186689
C -0.004506379 -0.02610882 -0.20411766  0.72301123 -0.65946096
D  0.480278227  0.87402135 -0.06425024  0.03175734  0.01681909
E  0.009675349  0.07161333  0.93713580 -0.04183150 -0.33882817
> plot(result)
> biplot(result)
```

De två plotten som producerades finns i Figur 2 och Figur 3.

(a) Vad visar Figur 2 om data? (1 poäng)

(b) Vad visar Figur 3 om data? (Formulera et par viktiga observationer) (2 poäng)

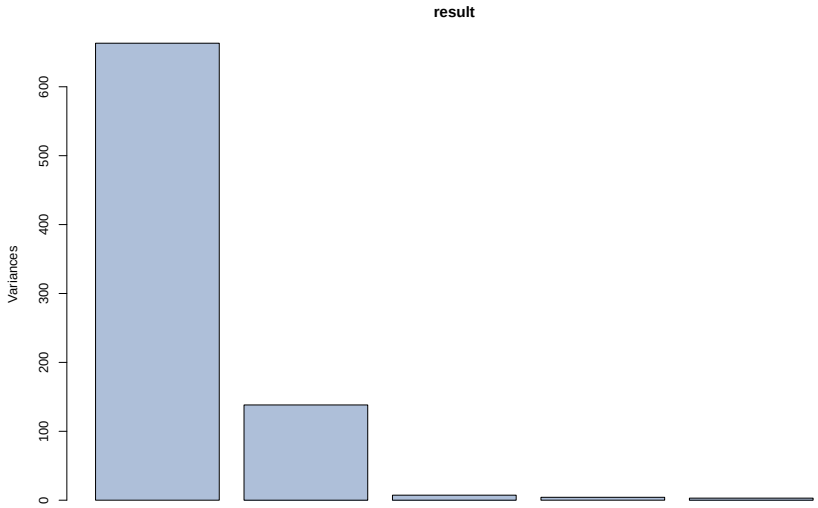

Figure 2: The result of the plot(result) command

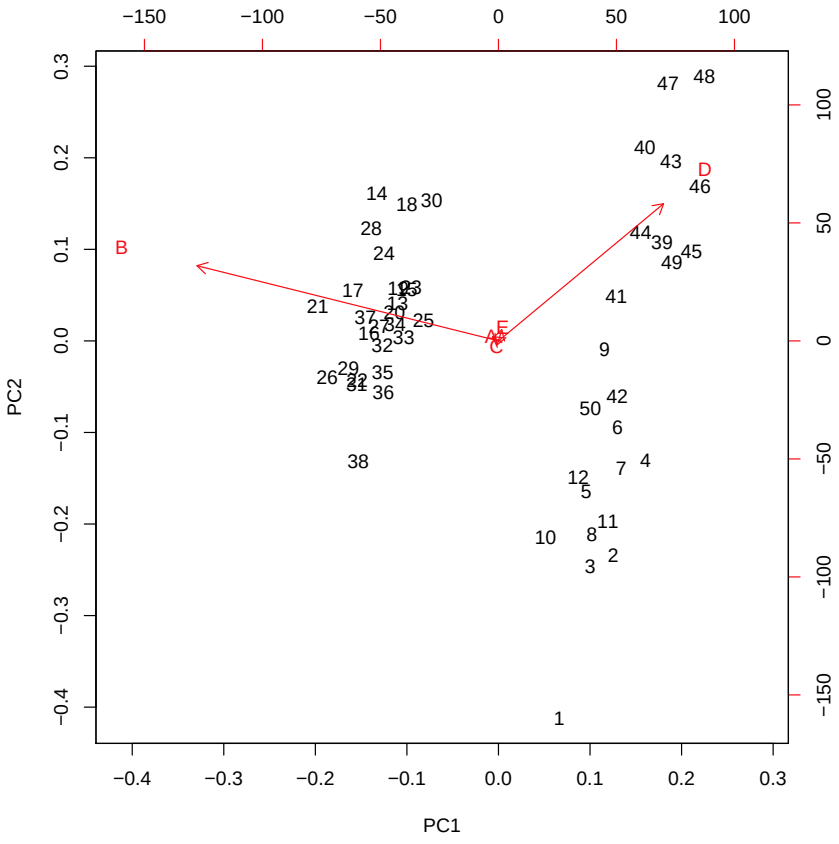

Figure 3: The result of the biplot(result) command