Statistische gegevensanalyse

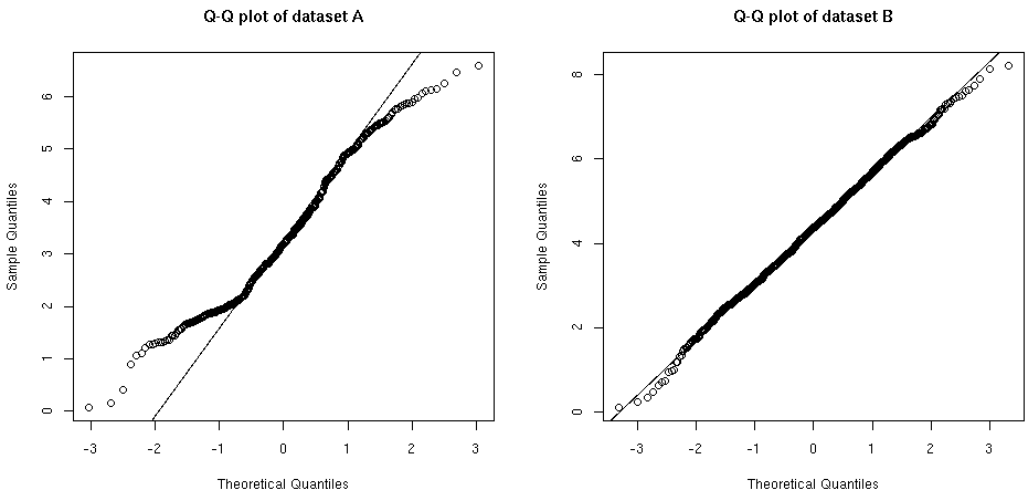
Bacheloropleiding in de Informatica

Examen – 21 juni 2016

- GEEN GSM, GEEN rekenmachine!
- Geef de nodige berekeningen, eindresultaten zijn onvoldoende.
- Verklaar alle overgangen en omkader het eindresultaat.
- Er wordt bij de quotering ook rekening gehouden met een duidelijke en logische opbouw van je antwoord.

Concepten uitleggen

1. Waarom hebben we een betrouwbaarheidsinterval nodig, indien we over een schatting beschikken zoals het steekproefgemiddelde of de mediaan?
2. We weten dat voor n normaal verdeelde observaties met gekende variantie σ^2 de lengte van het 95% betrouwbaarheidsinterval voor het populatiegemiddelde gegeven is door $L = 2 * 1.96 * \sigma / \sqrt{n}$. Leg de rollen/effecten van de steekproefgrootte n en van de variantie σ op de lengte van het interval uit.
3. Stel dat je een hypothese toets uitvoert op basis van data $x_1, \ldots, x_n$. Je doet een toets op het niveau 5%, en je vriend gebruikt dezelfde toets op het niveau 1%. Voor wie van jullie twee wordt de nulhypothese vroeger verworpen, en waarom? Gebruik een tekening om je antwoord duidelijk te maken.
4. Waar of niet? De p-waarde is de kans dat de nulhypothese waar is.
5. Waarom kan men de klassieke lineaire regressie niet gebruiken voor classificatie problemen?
6. Zie volgende QQ-plots. Wat kunnen we beslissen op basis van deze plots? De bedoeling is hier om te zien of dataset A en dataset B een normale verdeling volgen.



Oefeningen

1. In een museum, dat zaterdag open is van 9 tot 17u en zondag van 10 tot 16u, wordt elke dag het aantal bezoekers bijgehouden. Zo weet men dat het aantal elke dag Poisson-verdeeld is, op zaterdag met gemiddeld 2 bezoekers per uur en op zondag met gemiddeld 2,5 bezoekers per uur. De dagelijkse bezoekersaantallen zijn verder onafhankelijk van elkaar.
   Door een misverstand weet men enkel dat er vorig weekend in totaal 25 personen het museum bezocht hebben. Wat is de kans dat het er op zondag precies 10 waren? De formule voor de Poisson-verdeling is als volgt: $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$.
2. Beschouw de dataset `examen.txt`. Deze dataset bestaat uit twee variabelen, `Var1` en `ind`. De variabele `ind` verdeelt de gegevens in twee groepen.
   (a) Geef het gemiddelde, de mediaan, de variantie en de 35% en 65% percentielen voor `Var1`, zowel voor alle data als voor de twee groepen apart.
   (b) U wil nagaan of de data uit de twee groepen hetzelfde populatiegemiddelde hebben. Welke test zou u hiervoor gebruiken? Geef de nul- en de alternatieve hypothese.
   (c) Ga de voorwaarden voor de door u gekozen test na. Mag u de test toepassen?
   (d) Voer de door u gekozen test uit op het 5%-significantieniveau en interpreteer de resultaten.
3. In het midden van de jaren 1970 moest de universiteit van Berkeley zich verdedigen tegen beschuldigingen van discriminatie t.o.v. potentiële vrouwelijke studenten. De verdeling van de studenten die in 1973 een aanvraag indienden om aan deze universiteit te mogen studeren, zag er als volgt uit: 1158 mannelijke en 557 vrouwelijke studenten werden toegelaten, terwijl er 1493 mannen en 1278 vrouwen werden geweigerd.
   (a) Voer een eerste analyse uit die nagaat of er een verband bestaat tussen geslacht en de kans om toegelaten te worden. Geef duidelijk de hypotheses waaronder u werkt.
   (b) In onderstaande tabel vindt u de cijfers in meer detail terug, per departement. Bereken de toelatingspercentages voor de verschillende departementen. Liggen

Tabel 1: Toelatingscijfers UC Berkeley per departement

	Mannen		Vrouwen	
Departement	Toegelaten	Afgewezen	Toegelaten	Afgewezen
A	512	313	89	19
B	313	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317

   deze percentages in lijn met het resultaat van de vorige vraag? Verklaar.
4. Beschouw de dataset `ChickWeight`. Deze dataset is automatisch ingeladen in `R` en `RStudio`. U kan een beschrijving van de variabelen vinden in de dataset met behulp van het commando `?ChickWeight`.
   (a) Verwacht u een significant verschil in het gewicht tussen de verschillende groepen kuikens op dag 0? Stel dat u een significant verschil zou opmerken, wat zou dit kunnen betekenen?

(b) Voer een One-way ANOVA uit om de gemiddelde gewichten van de verschillende groepen kuikens te vergelijken op dag 0. Formuleer de nulhypothese en de alternatieve hypothese en ga de voorwaarden van de toets na.

(c) Voer een One-way ANOVA uit om de gemiddelde gewichten van de verschillende groepen kuikens te vergelijken op dag 21. Formuleer opnieuw de nulhypothese en de alternatieve hypothese en ga de voorwaarden van de toets na.

(d) Voer een post-hoc analyse uit waarbij u de Bonferroni-correctie toepast.

5. Beschouw de dataset `Titanic.txt`. Deze dataset heeft vier variabelen: `Class`, `Age`, `Sex` en `Survived`.

Tabel 2: Overzicht van de variabelen uit de dataset

`Class`	Eerste (1), tweede (2), derde (3) klasse en bemanning (0)
`Age`	Kind (0) of volwassene (1)
`Sex`	Man (0) of vrouw (1)
`Survived`	Gestorven (0) of overleefd (1)

(a) Voer een logistische regressie uit met de variabele `Survived` als afhankelijke variabele en de andere drie variabelen als onafhankelijke variabelen. Schrijf het logit-model uit en vul de gevonden waarden voor de coëfficiënten in. De algemene formule is $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$.

(b) Interpreteer de waarde van de coëfficiënt β bij de variabele `Age` in het logit-model. Interpreteer het effect van de variabele `Age` op de overlevingskans.

(c) Wat was de overlevingskans van een volwassen mannelijk bemanningslid op de Titanic volgens het gevonden model?