UNIVERSITEIT
GENT

Examen niet-parametrische statistische methoden 2018

Uw opdracht bestaat er in onderstaande vragen op te lossen en de antwoorden neer te schrijven in een beknopt en volledig rapport. De output van de statistische software kan als appendix toegevoegd worden. Breng op het examen het rapport met de appendices mee. Op het examen overlopen we uw rapport en indien er onduidelijkheden zijn, dan zal ik hierover vragen stellen. Tenslotte kunnen er ook enkele vragen gesteld worden om na te gaan of de relatie theorie-toepassing goed begrepen is. Onderstaande opdrachten zijn vrij summier neergeschreven en er is wat vrijheid in hoe je precies de analyses uitvoert en rapporteert. Hecht voldoende aandacht aan het visualiseren van de data en het neerschrijven van de conclusies. Tracht ook steeds zo nauwkeurig mogelijk de onderzoeksvraag te beantwoorden (dit heeft voornamelijk betrekking op de eerste twee vragen). Samen met de finale resultaten is het denkproces belangrijk. Geef daarom in het rapport beknopt weer waarom je bepaalde methodes hebt gekozen. Het rapport mag maximaal 5 pagina's lang zijn (zonder de appendices).
Gelieve het rapport te mailen naar Jan.DeNeve@UGent.be ten laatste op 22 augustus 2018 23u59 met al titel in de mail 'Examen UGain NPSM'. Het rapport moet een pdf zijn met naam 'VoornaamAchternaam.pdf' en gelieve samen met het rapport ook je R code door te sturen.

Veel succes!

1. Een bepaalde theorie met betrekking tot het geheugen veronderstelt dat de graad waarmee taalkundige gevens worden herinnerd, afhangt van het niveau waarop de gegevens worden verwerkt. Een onderzoeker heeft at random 30 jongere personen en 30 oudere personen onderverdeeld in 3 groepen.

    - De "Counting"groep werd gevraagd om een lijst van woorden te lezen en om de letters te tellen van elk woord. Dit is een laag verwerkingsniveau
    - De "Imagenery"groep werd gevraagd om zich elk woord levendig voor te stellen. Dit is een hoger verwerkingsniveau.

Aan de eerste 2 groepen werd niet meegedeeld dat het de bedoeling was om na te gaan hoeveel van de woorden ze zich zouden herinneren.

- Tot slot werd aan de "Intentional"groep gevraagd om elk woord te memoriseren en werd hen meegedeeld dat ze die later zouden moeten neerschrijven. Dit is het hoogste verwerkingsniveau.

Na de verwerkingsfase werd gevraagd aan de proefpersonen om zo veel mogelijk woorden neer te schrijven. De dataset `memory.rda` kan je terugvinden op Zephyr en bevat volgende variabelen:

- Age: Younger or Older
- Process (de graad van verwerking): Counting, Imagery of Intentional
- Words: Aantal woorden dat de proefpersonen kunnen reproduceren

De onderzoeksvraag is de volgende: "Is er een effect van de graad van verwerking op het aantal woorden dat de proefpersonen kunnen reproduceren en hangt dit effect mogelijks af van de leeftijd?"

(a) Gebruik een parametrisch lineair regressiemodel om de onderzoeksvraag te beantwoorden. Bespreek de assumpties onderliggend aan deze analyse.

(b) Voer een tweede analyse uit op basis van ranggebaseerde methodes. Bij deze methodes kan je niet expliciet testen voor interactie, maar je kan er wel impliciet mee rekening houden door 2 analyses te lopen: een analyse voor de jongere personen en een analyse voor de oudere personen.

(c) Bespreek en verklaar de verschillen tussen beide analyses (regressie versus ranggebaseerde methodes).

2. Onderzoekers wensen na te gaan indien er een lange termijn effect en een seizoenseffect is op de werkloosheidsgraad in de U.S. tussen 1970 en 1990; ze kunnen vooraf niet veronderstellen dat het mogelijke effect lineair is.

   De dataset `unemploymentUS.rda` kan op Zephyr worden teruggevonden en bevat de volgende variabelen:

   - Year: jaartal
   - Monthly: maandnummer
   - Rate: werkloosheidsgraad (%)

   Analyseer de data volgens een methode naar keuze.

3. De laatste opdracht bestaat uit het opzetten van een Monte-Carlo simulatie waarbij je de two-sample t-test evalueert onder de nulhypothese $H_0 : \mu_1 = \mu_2$. Je moet verschillende scenario's beschouwen:

- de data komen uit een normale verdeling versus de data komen uit een lognormale verdeling (ter inspiratie zie de R-code onderaan om te simuleren uit een lognormale).
- de standaarddeviaties zijn gelijk ($\sigma_1 = \sigma_2$) versus ze zijn verschillend ($\sigma_1 = 5 * \sigma_2$). Je bent vrij om σ_1 zelf vast te leggen (idem voor μ_1).
- de totale steekproefgrootte ($n_1 + n_2$) is gelijk aan 20 of 200.
- de data zijn gebalanceerd ($n_1 = n_2 = (n_1 + n_2)/2$) versus niet-gebalanceerd ($n_1 = (n_1 + n_2)/4$ en $n_2 = 3(n_1 + n_2)/4$).
- gebruik zowel de pooled-variance two-sample t-test als de Welch two-sample t-test (zie `?t.test` bij `var.equal` voor meer informatie).

Voor alle combinaties van bovenstaande scenario's (er zijn 32 scenario's in totaal), simuleer de type I fout voor $\alpha = 0.05$ op basis van 10000 Monte-Carlo simulaties. Rapporteer de resultaten in een overzichtelijke tabel en bespreek wanneer welke toets (bij benadering) de type I fout correct controleert en wanneer niet. Onderaan wat R code die je kan gebruiken om te simuleren uit een lognormale verdeling waarbij de standaardeviaties gelijk of verschillend kunnen zijn en waarvoor de gemiddeldes steeds gelijk zijn.

```
> simuleer.lognormale <- function(n1, n2, ratio = 1)
+ {
+   Y1 <- exp(rnorm(n1))-exp(.5) # gemiddelde = 0, zie wiki lognormale
+   Y2 <- exp(rnorm(n2))-exp(.5) # gemiddelde = 0
+   Y1 <- Y1*ratio # herschaal, ratio = 1 (gelijke varianties) of
+    5 (verschillende varianties)
+   return(c(Y1,Y2))
+ }
> # Eens snel via Monte-Carlo kijken of de code naar behoren werkt
> n1 <- 1000000 # een groot getal om de populatie te benaderen
> n2 <- 1000000
> Y <- simuleer.lognormale(n1, n2, ratio = 1)
> Y1 <- Y[1:n1]; Y2 <- Y[(n1+1):(n1+n2)]
> mean(Y1);mean(Y2) # gelijk aan elkaar (bij benadering)
[1] -0.0007310769
[1] -0.0004972798
> sd(Y1);sd(Y2) # ook gelijk
[1] 2.176685
[1] 2.159369
> Y <- simuleer.lognormale(n1, n2, ratio = 5)
> Y1 <- Y[1:n1]; Y2 <- Y[(n1+1):(n1+n2)]
> mean(Y1);mean(Y2) # gelijk
```

```
[1] 0.01549881
[1] -0.003645195
> sd(Y1);sd(Y2) # een factor 5 verschil
[1] 10.86477
[1] 2.164858
```