Academiejaar 2021–2022 – 2^e examenperiode	**Individuele opgave**
Departement: IT en Digitale Innovatie Opleiding, afstudeerrichting en jaar: Toegepaste informatica, 2TI Naam van het opleidingsonderdeel: Data Science & AI dOLOD/Deelexamen: Theorie+oefeningen Campus: Aalst, Schoonmeersen, TIAO, VC Lesgever(s): Sabine De Vreese, Stijn Lievens, Lieven Smits, Bert Van Vreckem	Examendatum: 2022-05-31 Aanvangsuur examen: 8:30
Voornaam en naam student: DEMUNTER Yoran	
Studentennummer: 202075141	
Lector bij wie de student de onderwijsactiviteit volgde:	Lesgroep v/d onderwijsactiviteit: PBA-TIN-TI/G2B4
Het behaalde resultaat op dit examen wordt herleid tot een cijfer op 20.	

Het examen is open boek. Tijdens het examen mogen volgende hulpmiddelen gebruikt worden:

- Eigen laptop met alle daarop aanwezige software zoals Python, Excel (of andere rekenbladsoftware), rekenmachine, ...
- Internetverbinding voor gebruik van Google Colab, raadplegen elektronisch cursusmateriaal via Chamilo of Github, opzoeken van informatie.
- Afgedrukt cursusmateriaal, bv. slides, eigen uitgewerkte oefeningen, Python-code, nota's, ...
- **Je mag geen enkele vorm van communicatie gebruiken tijdens de examens (chatten, mailen, Discord, Messenger, GSM, personen bij je in de buurt, ...).**

Algemene richtlijnen:

- Gebruik het bijgevoegde `.ipynb`-bestand (Jupyter Notebook) om je antwoorden op de vragen in te vullen, met eventueel de Python-code die je gebruikt hebt om het resultaat te bekomen.
- Je hernoemt het bestand naar **dsai-demunter-yoran.ipynb**
- Het moet mogelijk zijn om alle code in het notebook opnieuw uit te voeren zonder fouten.
- Wanneer de uitkomst een reëel getal is, rond dan af tot **exact drie cijfers na de komma**, behalve als dit **expliciet anders gevraagd wordt**. Er worden geen marges gerekend op de antwoorden.
- **Let op!** Geef altijd een duidelijk en expliciet antwoord op de (onderzoeks-)vraag in een Markdown-cel. Enkel de Python-code om het antwoord te bekomen of de uitvoer van de Python-code is onvoldoende. De Python-code (en de uitvoer ervan) is enkel ter info, en geldt niet als antwoord.
- Dien het ipynb-bestand in op exam.hogent.be.
- **Wie indient na de deadline krijgt Afwezig.** Wacht dus niet tot de laatste minuut om in te dienen. Dien eventueel tijdens het examen af en toe een onafgewerkte versie in.

Veel succes!

Vragen

1. **(2 pt)**

   Wat is het meetniveau van deze variabelen?

   (a) Het melkras van een koe (roodbont, Holstein, enz)

   (b) Het gewicht van een koe (in kg)

   (c) De jaarlijkse melkproductie van een koe (in l)

   (d) Levensfase van een koe (kalf, pink, vaars, koe)

2. **(2 pt)**

   Je herinnert je vast nog de vaccinatiecampagne tijdens de COVID-19 pandemie: alle burgers kregen de kans om zich te laten vaccineren, met ouderen en risico-patiënten eerst, en daarna volgens dalende leeftijd. Een nieuwszender wou in begin van deze campagne weten of de gemiddelde Vlaming tevreden was over de werking van het vaccinatiecentrum en vatte daartoe post aan de uitgang van verschillende vaccinatiecentra om 1000 respondenten te verzamelen.

   (a) Is dit een aselecte steekproef? Motiveer je antwoord.

   (b) Welk soort fout wordt hier gemaakt? Leg uit.

3. **(4 pt)**

   Beschouw volgende kansfunctie voor een toevalsveranderlijke X:

x	-3	-2	2	3
$f_X(x)$	1/16	z	3/16	2/16

   (a) Bepaal z zodat we te maken hebben met een geldige kansfunctie.

   (b) Geef μ_X (ook genoteerd als $E(X)$)

   (c) Bereken σ_X

   (d) Bepaal $P(X \geq -2)$

4. **(2 pt)**

   Patiënten die herstellen van een beroerte laten hun grijpkracht in elke hand laten meten om hun vooruitgang te volgen. Een bepaalde populatie van meer dan 100 mannelijke patiënten hebben grijpkracht in hun dominante handen met een gemiddelde en standaardafwijking van respectievelijk ongeveer 41 kg en 9 kg.

   Stel dat we aselecte steekproeven nemen van 4 mannelijke patiënten uit deze populatie en berekenen het steekproefgemiddelde $\overline{x}$ van elke groep patiënten. Wat zal de vorm zijn van de kansverdeling van $\overline{x}$?

   (a) Onbekend. We hebben niet genoeg informatie om de vorm te bepalen.

   (b) Scheef naar rechts.

   (c) Scheef naar links.

   (d) Benaderend normaal verdeeld.

5. **(2 pt)**

   Uit een normaal verdeelde populatie werd de hieronder gegeven steekproef getrokken. Bereken een 95%-betrouwbaarheidsinterval voor het populatiegemiddelde.

```
sample = np.array([
    5371, 5065, 5740, 5089, 4078, 4508, 4519, 4770, 5097, 4595, 5656,
    4443, 5762, 5050, 4238, 3485, 5056, 5390, 4825, 4908, 4023, 5026,
    3975, 4430, 5287
])
```

(a) Ondergrens

(b) Bovengrens

6. **(6 pt)**

   De steekproef die hieronder gegeven is, bevat gerelateerde meetresultaten van een variabele Count op twee tijdstippen (*time1* en *time2*), telkens op een overeenkomstige plaats. We willen weten of de metingen op het eerste tijdstip significant verschillend zijn van die op het tweede tijdstip.

   Gebruik een geschikte statistische toets (met significantieniveau 5%) om deze uitspraak te verifiëren.

```
my_sample = pd.DataFrame(data={
  'time1':
    [267, 252, 170, 275, 210, 179, 186, 198, 198, 249, 206, 221, 177, 161,
      251, 187, 199, 157, 184, 191, 133, 199, 187, 202, 196, 169,
      213, 287, 213, 172],
  'time2':
    [163, 227, 213, 248, 188, 187, 178, 218, 179, 276, 279, 183, 200, 219,
      132, 145, 258, 153, 197, 240, 214, 124, 203, 189, 195, 216,
      235, 167, 188, 138]
})
```

   (a) Welke toets moet je gebruiken om deze onderzoeksvraag te beantwoorden? Wees zo specifiek mogelijk!

   (b) Formuleer de nulhypothese en de alternatieve hypothese

   (c) Bereken de p-waarde

   (d) Trek een besluit op basis van de vorige stap en beantwoord de onderzoeksvraag.

   (e) Geef de mediaan voor de groep time1

   (f) Geef het bereik voor de groep time1

7. **(4 pt)**

   Een marketing-bureau wil via een enquête te weten komen of mensen een voorkeur hebben voor het meest dystopische bedrijf. Hieronder zijn de resultaten te vinden.

   Komt elke keuze ongeveer evenveel voor, of komt er uit de resultaten een duidelijke voorkeur naar voor? Gebruik een geschikte statistische toets (met significantieniveau 5%) om dit te onderzoeken.

```
df = pd.DataFrame(data={'Preference': [
  "Dharma", "Dharma", "LexCorp", "Cyberdyne", "Dharma", "Umbrella",
  "Umbrella", "LexCorp", "Umbrella", "Umbrella", "Tyrell",
  "Umbrella", "Umbrella", "Tyrell", "Umbrella", "Dharma", "LexCorp",
  "Tyrell", "Umbrella", "Umbrella", "Umbrella", "Umbrella", "Tyrell",
  "Tyrell", "Umbrella", "Umbrella", "LexCorp", "Umbrella", "Dharma",
  "Cyberdyne", "LexCorp", "Tyrell", "Cyberdyne", "Dharma", "Dharma",
```

```
  "Umbrella", "Umbrella", "LexCorp", "Dharma", "Umbrella",
  "Umbrella", "Cyberdyne", "Cyberdyne", "Umbrella", "Cyberdyne",
  "Umbrella", "LexCorp", "Tyrell", "LexCorp", "Cyberdyne",
  "Umbrella", "Tyrell", "Cyberdyne", "Dharma", "Umbrella", "Tyrell",
  "Dharma", "Umbrella", "Dharma", "Dharma"
]})
```

(a) Welke toets moet je gebruiken om deze onderzoeksvraag te beantwoorden? Wees zo specifiek mogelijk!

(b) Bereken de p-waarde

(c) Trek een besluit op basis van de vorige stap en beantwoord de onderzoeksvraag.

(d) Geef het gestandaardiseerde residu voor variabele Dharma.

(e) Kan je daaruit besluiten dat de waarde voor Dharma extreem over- of ondervertegenwoordigd zijn in de steekproef? Antwoord met een van volgende mogelijkheden (oververtegenwoordigd/ondervertegenwoordig/niet extreem).

8. **(6 pt)**

Beschouw de steekproef die hieronder gegeven is

```
sample_data = pd.DataFrame(data = {
  'x': [450, 507, 510, 553, 535, 601, 564, 550, 432, 527, 429, 482, 508,
        542, 440, 502, 441, 431, 543, 518, 391, 459, 483, 499,
        433, 539, 488, 592, 524, 514, 512, 483, 438, 522, 574,
        507, 437, 438, 627, 482, 578, 520, 500, 556, 547, 509,
        501, 524, 380, 479, 579, 510, 545, 495, 439, 501, 543,
        523, 605, 441, 554, 452, 481, 547, 398, 527, 546, 474,
        470, 569, 500, 568, 511, 450, 419, 500, 517, 583, 475,
        440],
  'y': [491, 493, 493, 481, 482, 469, 469, 476, 498, 493, 508, 489, 487,
        485, 504, 495, 501, 498, 489, 475, 503, 501, 495, 498,
        506, 484, 485, 465, 483, 493, 488, 491, 497, 481, 485,
        492, 498, 498, 471, 489, 477, 482, 494, 480, 477, 493,
        491, 483, 511, 497, 471, 486, 483, 495, 500, 484, 479,
        486, 468, 503, 482, 500, 488, 473, 513, 492, 479, 504,
        494, 478, 490, 482, 486, 495, 499, 490, 483, 476, 495,
        508]})
```

(a) Bereken de richtingscoëfficiënt van de regressierechte

(b) Bereken het snijpunt van de regressierechte met de y-as

(c) Bereken de correlatiecoëfficiënt (symbool + waarde)

(d) Geef een interpretatie voor de waarde van de correlatiecoëfficiënt

(e) Bereken de determinatiecoëfficiënt (symbool + waarde)

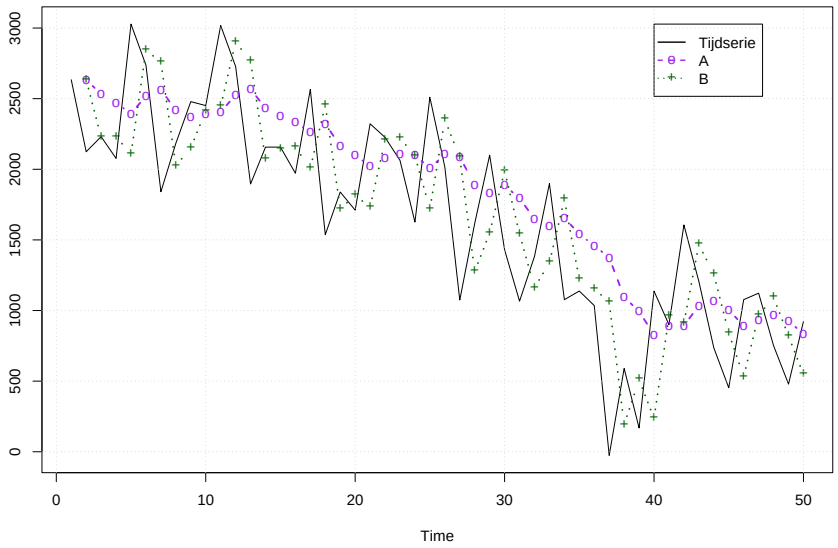
(f) Geef een interpretatie voor de waarde van de determinatiecoëfficiënt

9. **(2 pt)**

In Figuur 1 vind je een grafiek van een tijdserie (in het zwart) met twee vormen van voortschrijdend gemiddelde in kleur getoond.

Wat zijn de lijnen A en B precies? Kies uit de mogelijkheden hieronder:

- Eenvoudig voortschrijdend gemiddelde met periode 10



Figuur 1: Tijdseriedata (in het zwart) met twee vormen van voortschrijdend gemiddelde (in kleur).

- Eenvoudig voortschrijdend gemiddelde met periode 15
- Eenvoudig voortschrijdend gemiddelde met periode 20
- Enkelvoudige exponentiële afvlakking met α = 0.2
- Enkelvoudige exponentiële afvlakking met α = 0.8

Ter info is de definitie van de tijdserie hieronder gegeven, maar je hebt deze in principe niet nodig om de vraag te kunnen beantwoorden:

```
df = pd.DataFrame(data={'observations':
    [2634, 2124, 2231, 2077, 3028, 2737, 1841, 2184, 2479, 2451, 3018,
      2728, 1896, 2157, 2157, 1972, 2566, 1537, 1839, 1711, 2322,
      2226, 2063, 1626, 2510, 2016, 1074, 1617, 2100, 1430, 1066,
      1383, 1899, 1077, 1138, 1035, -26, 590, 168, 1137, 900, 1606,
      1206, 736, 453, 1077, 1123, 751, 479, 923]
})
```

Vermeld expliciet welke lijn welk voortschrijdend gemiddelde voorstelt.

(a) Lijn A: . . .

(b) Lijn B: . . .