# Opdrachten Statistiek 3 BIN (22 juni 2020)

Tijdens het maken van deze opdrachten mag je gebruik maken van alle informatie die op Blackboard en/of internet staat. Je mag alleen niet onderling overleggen, code en/of antwoorden uitwisselen. Ieder van jullie krijgt *individuele* datasets, deze staan op de BB toetspagina als zip-file achter je naam. Lever je werk in op BB als **R Markdown document** èn maak van je uitwerkingen een **html-file** waarin zowel de R-code, de output van R en de grafieken zichtbaar zijn. **Upload je R Markdown en html-file in één zip-file** naar BB (BB accepteert namelijk geen html-files!). Maak ook een kort filmpje waarin je je resultaten toelicht. Stuur dit filmpje per WeTransfer naar m.e.f.apol@pl.hanze.nl.

Cijfer = 1 + 9 * totaal / 175

Heel veel succes! Emile Apol

### Opgave 1 – Sterftecijfer Covid-19 voor twee bevolkingsgroepen in de VS

Het SARS-CoV-2 virus woedt nog steeds rond. Onder andere in de VS maakt het veel slachtoffers. Als snel kwam in het nieuws dat Afro-Americans onevenredig meer werden getroffen door Corona-19 dan blanken (Caucasians). Voor een aantal staten in de VS is het corona sterftecijfer per $100\ 000$ inwoners voor zowel Afro-Americans als Caucasians bepaald. De gegevens staan in de file `Opgave_1.txt`. Dit is een tab-separated ascii file met header.

**a.** [5 pt] Lees de data in, en maak in R een goede boxplot van het corona sterftecijfer als functie van bevolkingsgroep. Let op de as-titels en eenheden!

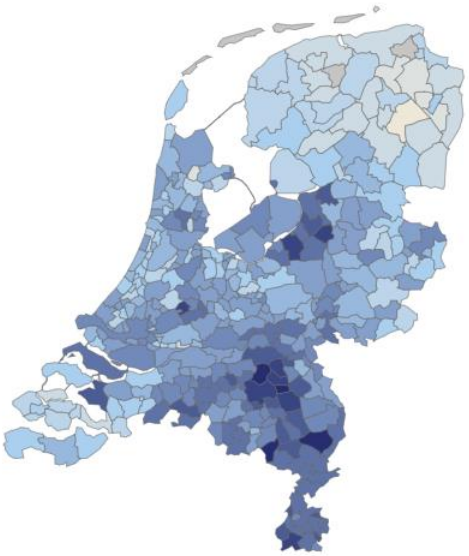
**b.** [5 pt] Leg uit welke statistische toets geschikt is om te onderzoeken of er een verschil is in corona sterftecijfer tussen Afro-Americans en Caucasians.

**c.** [5 pt] Is er een verschil in corona sterftecijfer tussen Afro-Americans en Caucasians? Toets met $\alpha = 0.05$.

**d.** [5 pt] Wat is de effectsterkte? Is het verschil in corona sterftecijfer tussen Afro-Americans en Caucasians een klein, matig of groot verschil?

**Opgave 2 – Ziekenhuisopnames Covid-19 patiënten in Nederland**

Ook Nederland is getroffen door de Covid-19 uitbraak. Een deel van de geinfecteerde personen moet worden opgenomen voor behandeling in het ziekenhuis. Gelukkig valt het aantal opnames in het noorden van ons land mee, maar er wonen hier ook minder mensen. In de file `Opgave_2.txt` staat per regio (Noord, Zuid, Oost en West) het aantal corona ziekenhuisopnames per $100\,000$ inwoners. Dit is een tab-separated ascii file met header, in "wide" format, met voor elke provincie een waarde.



**a.** [5 pt] Lees de data in R in, en zet het dataframe om naar een dataframe in "long" format.

**b.** [5 pt] Maak een goede boxplot van opnames per $100\,000$ als functie van regio. Let op de as-titels en eenheden!

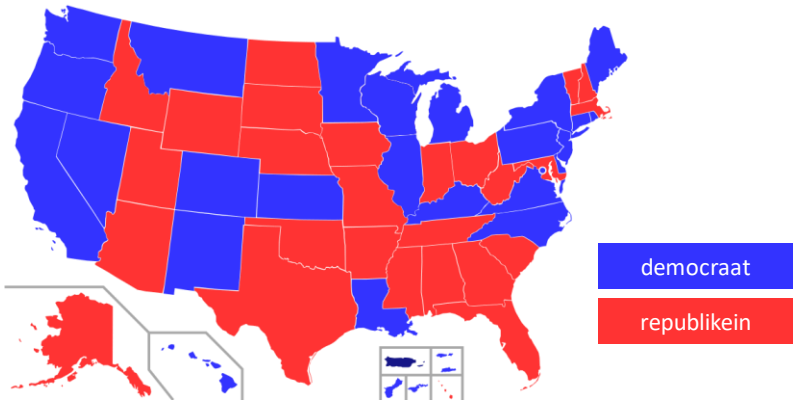
**c.** [5 pt] Welke statistische analyse is geschikt om te onderzoeken of het aantal corona ziekenhuisopnames per $100\,000$ inwoners verschilt per regio? Leg kort uit.

**d.** [5 pt] Is er een verschil in het aantal corona ziekenhuisopnames per $100\,000$ inwoners tussen de regio's? Toets met $\alpha = 0.05$.

**e.** [5 pt] Hoe groot is de effectsterkte? Heeft de regio een klein, matig of sterk effect op het aantal corona ziekenhuisopnames per $100\,000$ inwoners?

**f.** [5 pt] Voer een geschikte *post-hoc* test uit om te zien welke regio's hetzelfde aantal opnames per $100\,000$ inwoners geven. Toets met $\alpha = 0.05$.

**Opgave 3 – Sterftecijfer Covid-19 in de VS: republikeinen vs. democraten**




Het SARS-CoV-2 virus woedt nog steeds rond. Onder andere in de VS maakt het veel slachtoffers. De republikeinse regering o.l.v. Donald Trump ontkende eerst het gevaar van het virus. Democratische gouverneurs van verschillende staten kondigden echter al snel allerlei maatregelen aan om verspreiding van het virus te stoppen. Republikeinse gouverneurs volgden pas later maar beëindigden de maatregelen ook weer eerder. Democratische staten zijn vooral de stedelijke gebieden aan de oost- en westkust, republikeinse staten zijn vooral het dunbevolktere platteland in het midden. Voor 26 van de 51 random gekozen staten in de VS is het corona sterftecijfer van *blanken* per $100\,000$ inwoners bepaald (peildatum 9 juni 2020), alsmede de partij van de gouverneur: D = democraat of R = republikein. De gegevens staan in de file `Opgave_3.txt`. Dit is een tab-separated ascii file met header.

**a.** [5 pt] Lees de data in, en maak in R een goede boxplot van het corona sterftecijfer van blanken als functie van de partij van de gouverneur. Let op de as-titels en eenheden!

**b.** [5 pt] Leg uit welke statistische toets geschikt is om te onderzoeken of er een verschil is in corona sterftecijfer tussen democratische en republikeinse staten.

**c.** [5 pt] Is er een verschil in corona sterftecijfer tussen democratische en republikeinse staten? Toets met $\alpha = 0.05$.

**d.** [5 pt] Wat is de effectsterkte? Is het verschil in corona sterftecijfer tussen democratische en republikeinse staten een klein, matig of groot verschil?

**e.** [5 pt] Toont deze analyse *eenduidig* aan dat er een significant verschil zit tussen de resultaten van de democratische en de republikeinse aanpak van de coronacrisis? Licht je antwoord kort toe, en geef ook een mogelijke andere verklaring voor de eventueel gevonden verschillen in sterftecijfer.

## Opgave 4 – Corona sterftecijfer in de VS: effect van bevolkingsdichtheid?

Net als vele landen wordt de VS hard getroffen door de corona uitbraak. Aangezien het virus zich van persoon naar persoon verspreidt kun je je afvragen of het corona sterftecijfer (per $100\,000$ inwoners) van een staat binnen de VS afhankelijk is van de bevolkingsdichtheid (in aantal inwoners per km$^2$). De gegevens voor 25 van de 52 staten (op peildatum 9 juni 2020) staan in de file `Opgave_4.txt`. Dit is een tab-separated ascii file met header.

**a.** [5 pt] Lees de data in R in, en maak een goede scatterplot van het corona sterftecijfer als functie van de bevolkingsdichtheid. Let op de as-titels en eenheden!

Om te onderzoeken of er een (statistisch) verband is tussen sterftecijfer en bevolkingsdichtheid kun je in R lineaire *regressie* uitvoeren met het commando `summary(lm(y ~ x))` waarbij `y` het sterftecijfer is, en `x` de bevolkingsdichtheid. Je fit dan een lijn $y = a + b \cdot x$ door de grafiek; $a$ een $b$ zijn de coëfficiënten van het regressiemodel.

**b.** [5 pt] Voer lineaire regressie uit op de data. Wat is de waarde ("estimate") van de as-afsnede ("intercept") $a$, en wat is de waarde van de helling $b$?

De uitvoer van lineaire regressie geeft meer dan alleen maar de coëfficiënten $a$ en $b$: het geeft ook de standaardfouten in $a$ en $b$ ("Std. Error") en voert per coëfficiënt ook een $t$-toets uit met als hypotheses:

- $H_0$: de parameter is $0$;
- $H_1$: de parameter is niet $0$

De uitvoer van lineaire regressie geeft de $t$-waarde van deze toets ("t value") en de $p$-waarde van de toets ("Pr(>|t|)").

**c.** [5 pt] Is de waarde van de helling $b$ significant anders dan $0$? Toets met $\alpha = 0.05$.

**d.** [5 pt] Is er een significant verband tussen sterftecijfer en bevolkingsdichtheid?

De effectsterkte van regressie wordt gegeven door de waarde van $R^2$ ("Multiple R-squared"): dit is de fractie van de variatie in de $y$-waarden die wordt "verklaard" door het regressiemodel (dus: de lijn). De interpretatie is gelijk aan die van $\eta^2$ bij een 1-way ANOVA.

**e.** [5 pt] Wat is de effectsterkte van de bevolkingsdichtheid op het corona sterftecijfer? Heeft de bevolkingsdichtheid een klein, matig of sterk effect op het sterftecijfer?

**f.** [5 pt] Plot de gefitte regressielijn $y = a + b \cdot x$ in de grafiek van opgave **a.** Hint: gebruik de functie `abline`.

**g**. [5 pt] Wat zouden andere verklarende factoren voor het sterftecijfer kunnen zijn? Noem twee suggesties.

**Opgave 5 – Clustering van landen op basis van corona en luchtkwaliteit**

In Nederland zijn de gevolgen van het coronavirus in het zuiden het ergst, met name in Brabant en Limburg zijn veel (dodelijke) slachtoffers gevallen. In deze provincies is ook een hoge concentratie aan intensieve veehouderijen met varkens, koeien en kippen. Deze intensieve veehouderij zorgt voor een hoge concentratie fijnstof in de lucht. Ook industrie, verkeer, elektriciteitscentrales en branden in bossen en op akkers zorgen voor veel fijnstof. De concentratie fijnstof wordt uitgedrukt als $PM_{2.5}$ (in $\mu$g/m$^3$), d.w.z. de concentratie deeltjes met diameter kleiner dan $2.5$ $\mu$m. Het Zwitserse IQAir heeft gegevens verzameld van de luchtkwaliteit (= fijnstofconcentratie) van heel veel landen.

De Johns Hopkins Universiteit verzamelt wereldwijd de gegevens van het aantal Covid-19 besmettingen (per $100\,000$ inwoners), het aantal herstelde Covid-19 patiënten (per $100\,000$ inwoners) en het aantal overleden Covid-19 patiënten (per $100\,000$ inwoners).

In de file In de file `Opgave_5.txt` staat deze informatie (op peildatum 19 juni 2020) voor een grote groep landen. Dit is een tab-separated ascii file met header. `Confirmed`, `Recovered` en `Dead` zijn de aantallen besmettingen, herstellingen en doden per $100\,000$ inwoners per land, en `PM2.5` is de fijnstofconcentratie (in $\mu$g/m$^3$).

**a.** [5 pt] Lees deze file in R in als dataframe. Wat is het gemiddelde aantal besmettingen, herstelden en doden per $100\,000$ inwoners? En wat is de gemiddelde fijnstofconcentratie?

**b.** [5 pt] Voer hiërarchische clustering uit op de landen op basis van de kolommen `Confirmed`, `Recovered` en `Dead`. Gebruik euclidische afstanden en "average" linkage. Maak een dendrogram van het resultaat.

**c.** [5 pt] Splits de cluster in 2 subclusters m.b.v. de functie `cutree`. Welke landen horen bij subcluster 1 en welke landen horen bij subcluster 2?

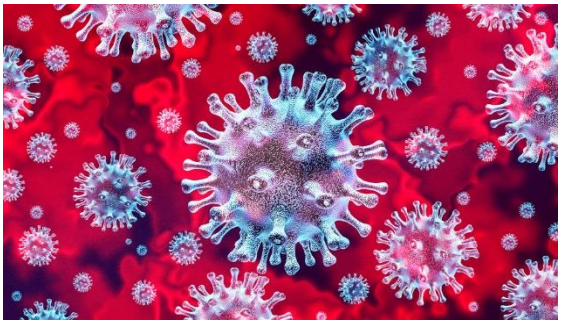
**d.** [5 pt] Toets of er een significant verschil zit in fijnstofconcentratie tussen de landen van subcluster 1 en van subcluster 2. Toets met $\alpha = 0.05$.

**e**. [5 pt] Voer $k$-means clustering uit op de landen op basis van de kolommen `Confirmed`, `Recovered` en `Dead`. Gebruik $k = 2$ clusters en gebruik $5$ random startwaarden voor de clustercentra. Maak m.b.v. de speciale functie `makeKMeansDendrogram` (zie map "R functies" op Blackboard) een dendrogram van het resultaat.

**f**. [5 pt] Zijn de twee $k$-means clustering subclusters hetzelfde als de twee hierarchische subclusters bij **c.**?

## Opgave 6 – Analyse op DEG's (1)

Om het effect van het Covid-19 coronavirus op de genexpressie van patiënten te onderzoeken wordt er van 6 personen die positief getest zijn op Covid-19 en van 6 personen die recentelijk negatief zijn getest met behulp van een 2-channel microarray de genexpressie van $G = 100$ genen bepaald. De resultaten staan in de file `Opgave_6.txt`. Dit is een tab-separated ascii file met header.



Per microarray geven het rode (R) en groene (G) signaal de log-getransformeerde en background-gecorrigeerde intensiteiten van de geïnfecteerde en niet-geïnfecteerde personen weer. Signaal "R.3" is bijvoorbeeld de log intensiteit van een geïnfecteerd persoon gemeten met microarray 3.

**a.** [5 pt] Lees de microarray data in R in als dataframe. Bekijk de structuur van de data.

**b.** [5 pt] Welke statistische toets is geschikt om direct dit dataframe te analyseren op de aanwezigheid van DEG's (differentially expressed genes)?

**c.** [5 pt] Maak een eigen functie die de toets van **b.** uitvoert per regel van het dataframe met microarray data.

**d.** [5 pt] Voer deze functie uit per gen (= regel van het dataframe). Hoeveel en welke genen komen *zonder* multiple testing correctie differentieel tot expressie (d.w.z. zijn DEG's)? Toets met $\alpha = 0.05$.

**e.** [5 pt] Pas nu FDR multiple testing correctie toe op de $p$-waarden. Hoeveel en welke genen komen na correctie differentieel tot expressie (d.w.z. zijn werkelijk DEG's)? Toets met $\alpha = 0.05$.

**f**. [5 pt] Hoeveel vals positieve uitslagen verwacht je zonder multiple testing correctie?

**g**. [5 pt] Maak van de bij **e.** gevonden DEG's een heatmap. Cluster zowel de genen als de samples o.b.v. euclidische afstand en average linkage. Gebruik een geschikt kleurenpalet.

**Einde van de opdrachten!**

Maak van je R Markdown file ook een html-bestand (via Knit), en maak een kort filmpje van een paar minuten waarin je je resultaten toelicht. Upload je R Markdown en html-bestand samen als zip-file naar Blackboard. Belangrijk: BB ondersteunt het uploaden van html-files niet, daarom graag zippen! Stuur je korte filmpje met toelichting via WeTransfer naar m.e.f.apol@pl.hanze.nl.