Computerpracticum
Grieks-Latijns vierkant en ANOVA

Inleiding
In dit (computer-)practicum gaan we een experimenteel ontwerp maken met een Grieks-Latijns vierkant. We doen vervolgens (gesimuleerd) het onderzoek in Excel. Daarna lezen we de gegevens in, in R en analyseren ze.
Onder het kopje "Opdracht" staat alleen *wat* je moet doen. Op zoek naar aanwijzingen voor wat er in R moet gebeuren? Die staan aan het eind van dit document, onder "Tips".

Casus
Ahmed is eigenaar van een Marokkaanse eet- en drinkgelegenheid, en serveert daar onder andere nana. Dat is een warme drank die wordt gemaakt met munt, groene thee en suiker. Er wordt een specifieke Chinese variant van groene thee gebruikt, en een muntvariëteit die in Marokko speciaal voor het maken van nana wordt geteeld. Ahmed wil echter experimenteren om na te gaan of hij een andere lekkere of gezondere variant kan maken. Hij wil verschillende soorten munt, verschillende soorten groene thee, en verschillende zoetmiddelen proberen, en ook experimenteren met de hoeveelheid munt.

De volgende ingrediënten zijn geselecteerd om uit te proberen.
- Voor de munt: mentha spicata nana (een variëteit van groene munt, Spic), mentha aquatica (watermunt, Aqua), mentha suaveolens (witte munt, Sua), en mentha piperita (pepermunt, Pip).
- Voor de groene thee: gunpowder (GP), mao feng (MF), sencha, en pu-erh (PE).
- Voor de zoetmiddelen: suiker, honing, aspartaam, en stevia.
- Voor de hoeveelheid munt: weinig, normaal, veel, en heel veel.

(Er zijn dus vier factoren, met elk vier mogelijke waarden.)

Ahmed is hoofdzakelijk geïnteresseerd in de hoofdeffecten van de verschillende ingrediënten; interacties zijn goed om te weten, voor zover mogelijk. Hij heeft er een avond voor uitgetrokken om verschillende recepten te proberen, zelf te proeven, en te waarderen met een rapportcijfer. Er is een belangrijke beperking: hij heeft tijd voor het proberen van maximaal 20 recepten. (En het onderzoek wordt dus in enkelvoud uitgevoerd.)

Opdracht

a) Maak het bij deze situatie passende vierkant waarin je de uit te voeren experimenten en instellingen aangeeft.
   Tip: Probeer niet "zomaar even" de lijst met experimenten te maken. Dat lukt vast niet. Kies welke variabele in je Grieks-Latijnse vierkant bij de rijen hoort, welke bij de kolommen, welke bij de Latijnse letters en welke bij de Griekse letter. Kies ook welke waarden van variabelen waarbij horen, en maak daarvan voor jezelf een lijstje. (Bijv. $\alpha = \ldots; \beta = \ldots$; et cetera.) Gebruik de structuur van een Grieks-Latijns vierkant zoals je die ergens kunt vinden, bijvoorbeeld in het dictaat.
b) Vertaal dit naar de lijst met uit te proberen recepten in het Excel sheet. (Invullen in de ruimte die daarvoor klaarstaat; de meetresultaten verschijnen vanzelf.)
c) Analyseer deze resultaten in R om na te gaan welke van de vier variabelen effect hebben op het resultaat.

d) Controleer je resultaat door R de relevante plaatjes te laten maken. Maak plaatjes van de effecten. (Alleen hoofdeffecten, geen interacties.) Daarvoor kun je het package `effects` installeren, om vervolgens met

```
plot( allEffects( L ) )
```

of iets dergelijks plaatjes bij je analyse te krijgen.

e) Geef, op basis van dit resultaat, het "ideale" recept. (Dus het recept dat leidt tot de hoogste score.)

f) Wat is, op basis van deze gegevens, je schatting van het verschil (in resultaat) tussen Gunpowder en Pu erh? (Dat kun je uit de getallen van de statistische uitvoer afleiden. Merk op dat onder *Estimate* de geschatte effecten staan.)

Tips

Gegevens inlezen in R. Installeer (als dat nog niet geïnstalleerd is) het R-package `readxl`. Dat package bevat de functie `read_excel`, waarmee je gegevens uit een Excelsheet kunt lezen:

`read_excel(`pad/bestand`, sheet =` sheet`, range =` bereik`, col_names = TRUE )`

Hierin moet je, uiteraard, voor *pad*, *bestand* en *sheet* de juiste waarden invullen, tussen aanhalingstekens, dus bijvoorbeeld

```
"d:/temp/data.xlsx"
```

of zoiets. Als je de working directory hebt ingesteld op de folder waarin de data staan, hoef je alleen de naam van het bestand op te geven, en kun je de folder weglaten.

Verder is het *bereik* het gebied in Excel waar de in te lezen gegevens staan, bijvoorbeeld `"B3:F8"`. (De hele rechthoek van B3 linksboven tot en met F8 rechtsonder wordt gelezen.) Met `col_names = TRUE` geef je aan dat de eerste rij geen gegevens bevat, maar de namen van de kolommen. (Laat je dit weg, dan wordt het waarschijnlijk goed geraden; het staat er hier even voor de volledigheid bij.)

Stop de ingelezen data wel in een variabele, zodat je er vervolgstappen mee kunt doen. En kijk vooral na het inlezen even goed of de boel goed is ingelezen, voordat je verder gaat.

ANOVA in R. In paragraaf 7.7 van het dictaat is terug te vinden hoe je in R een regressie en/of variantie-analyse kunt doen. We gebruiken daarvoor de functies `as.factor()`, `aov()` en `lm()`. Met `as.factor()` geef je aan dat een variabele niet als getallen moet worden behandeld[1]. Met `aov()` kun je nagaan welke variabelen zoal statistisch significante effecten hebben. En met `lm()` kun je inzicht krijgen in wat die effecten dan inhouden. Vergeet ook de functie `summary` niet, waarmee je de uitgebreide uitvoer op het scherm kunt krijgen.

[1] Deze functie heb je nodig als een variabele wel uit getallen bestaat, maar zonder dat je met die getallen wilt gaan rekenen. Een typische situatie waarin je dit tegenkomt is wanneer een nominale variabele is gecodeerd als getallen, bijvoorbeeld leeftijdscategorie met 1 = kind, 2 = volwassene, 3 = oudere. Bestaat je variabele uit strings (bijv. "kind") dan "snappen" functies als `anova`, `aov` en `lm` ook zonder `as.factor` wel dat daarmee niet moet worden gerekend.