KU LEUVEN kulak

Bachelor wiskunde, chemie, fysica, informatica,
ingenieurs- en bio-ingenieurswetenschappen,
handelsingenieur.

Tevredenheid van Belgische volwassenen

Academiejaar 2023 – 2024 Project statistiek

1 Toelichting bij het projectwerk

Het projectwerk is een onderdeel van het examen statistiek, telt mee voor 4 van de 20 punten en wordt (behoudens uitzonderingen) per drie gemaakt. Iedereen werkt samen met studenten uit dezelfde examengroep (`Examengroepen.pdf`). Er hoeft uiteraard slechts één gezamenlijk script en rapport te worden ingediend.

Het is de bedoeling om de leerstof in de praktijk te gebruiken. Daarom wordt er gewerkt met een realistische dataset die moet worden geïmporteerd in `R`. Met behulp van onderstaande onderzoeksvragen en opdrachten worden dan gepaste analyses uitgevoerd en conclusies getrokken. De evaluatie gebeurt op basis van volgende onderdelen:

1. Een script `naam1_naam2_naam3.R` met alle gebruikte commando's.
2. Een verslag `naam1_naam2_naam3.pdf` van *maximaal* 4 pagina's tekst (exclusief figuren en tabellen).

Beide bestanden worden ingediend via `Toledo`, vóór het einde van de lesperiode.

2 Gegevens en onderzoeksvraag

Het projectwerk zal de score onderzoeken die Belgische volwassenen zichzelf geven voor hun persoonlijke geluk. Er wordt nagegaan of en met welke socio-economische factoren de geluksscore samenhangt en of deze daaruit kan worden voorspeld.

Er zal gebruik worden gemaakt van gegevens die in 2016 werden verzameld binnen het MEqIn project [1], gevoerd door onderzoekers van 4 Belgische universiteiten waaronder medewerkers van de Faculteit Economie en Bedrijfswetenschappen KU Leuven Kulak. Het centrale thema van dit project is het *equivalent inkomen*, een term die recent in de economische literatuur is geïntroduceerd en die rekening probeert te houden met meer dan enkel inkomen of private consumptie maar ook met andere factoren zoals gezondheid, geluk en de verdeling van welzijn binnen het gezin.

Het projectwerk is gebaseerd op een kleine selectie van de beschikbare veranderlijken en deelnemers uit het volledige onderzoek. De observaties in de aangeleverde dataset zijn volwassenen die in België wonen, telkens één per gezin. Als veranderlijken zijn er enkele individuele karakteristieken van de respondent (`ind_XYZ`), gegevens over het gezin (*household*, `hh_XYZ`), de gezondheid (`health_XYZ`) en het aantal uur dat de respondent aan vrije tijd kan besteden (`leis_time`). De variabelen worden opgesomd in Tabel 1, een uittreksel uit de dataset is te zien in Tabel 2.

3 Opdrachten

3.1 Data inlezen en manipuleren

Om deze opdracht zo realistisch mogelijk te maken, wordt gestart vanaf het document `people2324.dat`. Een eerste moeilijkheid is immers vaak om gegevens correct in het statistiekpakket in te lezen. Kijk daarom zorgvuldig na hoe de data is gestructureerd en importeer ze volgens de richtlijnen uit de handleiding en de helpbestanden van `R`.

Vervolgens wordt de dataset onderzoeksklaar gemaakt. Geef de categorische veranderlijken passende labels. Maak verder de nieuwe veranderlijke `hh_parent` die aangeeft of de respondent ouder is van minstens één kind jonger dan 18 jaar en ook de veranderlijke `hh_alone` die aangeeft of de respondent alleen woont, dus zonder andere al dan niet volwassen gezinsleden.

Dit deel van de analyse valt buiten de statistiek en hoort als dusdanig ook niet in het verslag. Introduceer in het verslag meteen de dataset zoals die hier is geconstrueerd.

Tabel 1: Veranderlijken in de dataset.

	Naam	Beschrijving
1	ind_ID	Rangnummer van de respondent
2	ind_gender	Geslacht van de respondent (man = 1, vrouw = 2)
3	ind_age	Leeftijd van de respondent in jaar
4	ind_edu	Opleidingsniveau van de respondent (minder dan SO = 1, diploma SO = 2, hoger diploma = 3)
5	ind_happy	Geluksniveau van de respondent (uiterst ongelukkig = 0 tot uiterst gelukkig = 100)
6	ind_atwork	Respondent heeft betaald werk (nee = 0, ja = 1)
7	ind_income	Persoonlijk netto inkomen van de respondent in euro per maand (inkomens, pensioenen, toelagen)
8	hh_pos	Rol van de respondent in het gezin: geen inwonende partner (1), samenwonend met partner (2) of volwassene die bij ouders inwoont (3); andere respondenten (NA) wonen bijvoorbeeld in bij schoonouders of bij kinderen en kleinkinderen
9	hh_nadult	Aantal leden van het gezin van minstens 18 jaar
10	hh_nchild	Aantal leden van het gezin jonger dan 18 jaar
11	hh_income	Volledig beschikbaar inkomen van het gezin in euro per maand (inkomens, pensioenen, toelagen, kindergeld, vastgoed – geen kapitaalopbrengsten)
12	health_fys	Fysieke gezondheid van de respondent (problematisch = 0 tot probleemvrij = 100)
13	health_emo	Emotionele gezondheid van de respondent (problematisch = 0 tot probleemvrij = 100)
14	leis_time	Tijd besteed door de respondent aan ontspanning in uur per week

Tabel 2: Uittreksel uit de dataset.

ind_ID	ind_gender	ind_age	ind_edu	ind_happy	ind_atwork	ind_income
10101	1	59	3	63	1	1860
10102	2	38	3	73	1	4600
10105	2	44	3	66	1	3600
10202	2	30	3	77	1	800
10203	1	60	3	80	1	1500
10205	1	50	3	75	1	3000
...						
25503	2	54	3	-	1	1490
...						
27010	2	65	1	-	0	87

ind_ID	hh_pos	hh_nadult	hh_nchild	hh_income	health_fys	health_emo	leis_time
10101	2	2	0	1860	3	37	30
10102	2	2	0	4600	100	88	18
10105	1	1	2	3929	83	80	10
10202	1	1	2	1057	83	67	4
10203	2	2	0	1500	100	88	3
10205	2	2	0	4100	100	77	20
...							
25503	1	2	1	1954	8	58	20
...							
27010	2	2	0	87	12	13	28

Tabel 3: Basisstatistieken.

Naam	Gemiddelde	Bereik	Algemene vorm
Lengte	$(12{,}3 \pm 0{,}4)$ m	$[5{,}6\,\text{m}, 27{,}8\,\text{m}]$	Eerder symmetrisch
Volume	$(12 \pm 3) \times 10\,\text{m}^3$	$[56\,\text{m}^3, 378\,\text{m}^3]$	Licht rechtsscheef
Massa	$(12{,}345 \pm 0{,}006) \times 10^6$ kg	$[5 \times 10^6\,\text{kg}, 678 \times 10^6\,\text{kg}]$	Lognormaal

Geslacht	Man	Vrouw
Aantal	23	44
Proportie	66%	34%

Studierichting	Biologie	Economie	Wiskunde	NA
Aantal	23	25	17	2
Proportie	35%	38%	26%	

3.2 Beschrijvende statistiek

Nu de dataset klaar is voor statistische analyses, kunnen ter verkenning voor elke veranderlijke gepaste datastatistieken worden berekend en grafieken worden gemaakt. Noteer voor elke veranderlijke of ze kwalitatief (nominaal of ordinaal) dan wel kwantitatief is (discreet of continu) en hoe de verdeling eruitziet: bekijk locatie en spreidingsparameter, wat is het bereik, is er symmetrie, komen er uitschieters voor? Bekijk waar relevant of een veranderlijke normaal verdeeld is en indien niet of een logaritmische transformatie de normaliteit verbetert. Vergelijk al even grafisch hoe de geluksscore die een respondent rapporteert, samenhangt met de andere veranderlijken.

Het is hier nog niet de bedoeling conclusies te trekken, maar om de gegevens beter te leren kennen. Het inlezen van data en verkennende beschrijvende statistieken horen doorgaans dan ook niet in het verslag, tenzij als illustratie bij het vervolg of indien er al opmerkelijke verbanden te zien zouden zijn. *Beperk verslaggeving van dit deel tot een overzichtstabel naar het model van Tabel 3, die basisstatistieken geeft bij elke numerieke veranderlijke en een frequentietabel bij de categorische.* Neem in deze tabel precies die veranderlijken op, die in het verslag aan bod komen. Vermeld dus *niet* de veranderlijken die nergens in je verslag aan bod komen en *wel* eventuele veranderlijken die pas later in het project worden aangemaakt (discretisaties, hercoderingen, ...).

3.3 Inferentiële statistiek

3.3.1 Kenmerken van de steekproef

Volgens StatBel, telde België op 1 januari 2016 precies 4 613 480 volwassen vrouwen en 4 368 849 volwassen mannen. De leeftijdsverdeling bij volwassenen was op het zelfde moment als volgt:

< 30	$[30, 40[$	$[40, 50[$	$[50, 60[$	$[60, 70[$	≥ 70
19%	16%	17%	18%	14%	16%

Nog volgens StatBel bedroeg het individuele netto belastbaar inkomen in dat jaar per inwoner gemiddeld 1485,33 euro.

Voer gepaste testen uit om na te gaan of de gegevens over de respondenten in de steekproef bovenstaande cijfers volgen. Deze analyses kunnen iets leren over de mate waarin het vervolg veralgemeenbaar is naar de rest van de populatie.

Uit het verslag moet duidelijk zijn welke testen werden uitgevoerd, hoe de voorwaarden werden nagegaan, wat de relevante statistieken zijn en welke conclusie er wordt getrokken. Illustreer significante resultaten.

3.3.2 Gemiddelde gelukscore

Verschilt de score voor geluk naargelang het geslacht van de respondent? Naargelang men betaald werk uitvoert? Is er een verschil tussen ouders met kinderen jonger dan 18 jaar naargelang ze een inwonende partner hebben?

Zet alle relevante statistieken samen in een overzichtelijke tabel en vat de conclusies bondig samen. Voorzie telkens een verduidelijkende grafiek.

Uit het verslag moet duidelijk zijn welke testen werden uitgevoerd, hoe de voorwaarden werden nagegaan, wat de relevante statistieken zijn en welke conclusie er wordt getrokken. Illustreer significante resultaten.

3.3.3 Associatie met de verschillende veranderlijken

Ga na of er afhankelijkheid is tussen de geluksscore enerzijds en de (niet-binaire) veranderlijken anderzijds.

Maak opnieuw een overzichtelijke tabel en vat bondig samen welke inzichten zijn verkregen. Voorzie telkens een verduidelijkende grafiek. Deze analyse geeft een idee van welke veranderlijken een rol kunnen spelen in het verklaren van de geluksscore.

Maak (een) overzichtelijke tabel(len) van deze resultaten. Vat bondig samen welke inzichten zijn verkregen. Voorzie verduidelijkende grafieken.

3.3.4 Verklaren van de geluksscore

Maak eerst twee verschillende eenvoudige regressiemodellen voor de geluksscore: het eerste in functie van het totale beschikbare inkomen van het gezin en het tweede in functie van het tiendelige logaritme van dat inkomen. Het enige doel van deze modellen is om één duidelijke figuur te maken met daarop de geluksscore in functie van het gezinsinkomen samen met beide eenvoudige regressiemodellen en de betrouwbaarheids- en predictiebanden. *Neem deze grafiek ter illustratie op in het verslag en leg aan de hand hiervan uit of de logaritmische transformatie noodzakelijk is.*

Bouw een meervoudig regressiemodel voor het verklaren van de geluksscore, door alle andere numerieke veranderlijken te gebruiken en achterwaartse regressie toe te passen. Ga de modelveronderstellingen na en gebruik waar nodig logaritmische transformaties van de gebruikte veranderlijken.

Ga vervolgens formeel na of het zin heeft afzonderlijke vergelijkingen te hanteren naargelang de respondent een man of een vrouw is.

Voor respondenten 25503 en 27010 ontbreekt de geluksscore (zie Tabel 2). Maak voor deze twee respondenten een voorspelling van de geluksscore met behulp van het gevonden model en bespreek de betekenis en kwaliteit van de bekomen voorspellingen en bijhorende intervallen. Duid deze voorspellingen ook aan op de eerder gemaakte grafiek met eenvoudige regressiemodellen.

Het verslag moet voldoende informatie bevatten voor de lezer om het bekomen model te reconstrueren. Geef de bekomen vergelijking(en) voor de geluksscore zo goed mogelijk. Bespreek de coëfficiënten en de kwaliteit van dit model. Voorzie diagnostische plots. Bespreek ook welke tekortkomingen het model eventueel nog heeft en hoe je het zou kunnen verbeteren.

4 Instructies

Script. Bundel alle commando's in een script. Zorg dat het script correct werkt op basis van het originele databestand. Verwijder alle overbodige lijnen en voeg zeer summier wat commentaar toe aan elke stap, zodat de commando's bij elk onderdeel vlot terug te vinden zijn. Het is niet nodig om in het verslag uit te weiden over technische details van `R`, over moeilijkheden bij het importeren van de data of over veranderlijken die je verder niet meer gebruikt.

Neem van de uitvoer van het script enkel die statistieken en grafieken in je verslag over die werkelijk relevant zijn voor de opbouw van het verhaal. In het bijzonder is het vaak nuttig om een illustratie te voorzien bij significante resultaten of een weerhouden regressiemodel. Noteer alle statistieken met de juiste eenheid en een gepast aantal beduidende cijfers zoals uitgelegd in Tabel 4. Zorg er voor dat je grafieken duidelijk leesbaar zijn en voorzien van titel, as-titels en eenheden. Gebruik vectorafbeeldingen (`.pdf`) in plaats van bitmaps (`.png`, `.jpg`) en zeker geen screenshots. Verwijs minstens één keer vanuit de tekst naar elke figuur.

Verslag. Maak van het rapport een degelijk wetenschappelijk verslag. Het moet een doorlopende tekst zijn, die los te lezen is van de opgave en begrijpelijk is voor een buitenstaander met dezelfde kennis van statistiek als jijzelf. Volg hoe dan ook de richtlijnen die er in jouw studierichting worden gehanteerd voor het schrijven van wetenschappelijke teksten. Via `Toledo` zal een sjabloon worden verspreid voor dit rapport met daarin een al uitgeschreven inleiding en onderstaande structuur. Ter illustratie zal ook een voorbeeldrapport worden verspreid dat volgens deze structuur is uitgewerkt.

Inleiding. Elk wetenschappelijk verslag begint met een korte introductie van het probleem, de dataset, de relevante veranderlijken en de onderzoeksvraag, vaak opgesplitst in meerdere hypothesen.

Methode. Vervolgens wordt per onderzoekshypothese uitgelegd welke statistische analyse zal worden gebruikt en hoe de voorwaarden worden nagegaan. Deze sectie kan in principe grotendeels voor het uitvoeren van de analyses worden geschreven. *Deze sectie beslaat maximaal 1 pagina.*

Resultaten. Een volgende sectie bevat alle numerieke resultaten van deze analyses, zo veel mogelijk in overzichtelijke tabellen. Dit deel bevat enkel objectieve gegevens, zoals statistieken en bekomen p-waarden, nog zonder interpretaties. *Deze sectie beslaat maximaal 1 pagina.*

Discussie. De interpretatie van de hypothesetesten en de verklaring van de statistische analyses horen in de discussiesectie. *Deze sectie beslaat maximaal 1 pagina.*

Besluit. Het besluit bestaat uit een compact en concreet antwoord op de onderzoeksvraag, geen herhaling van de resultaten maar een overkoepelende beschouwing bij de verschillende onderzochte hypotheses. Introduceer hier geen nieuwe elementen, maar probeer een zo algemeen mogelijke uitspraak te doen over wat de analyse je globaal heeft geleerd.

Bij elke opdracht staan specifieke aanwijzingen voor wat er precies in het verslag hoort (cursieve tekst). Volg deze en controleer na afloop of alle gevraagde elementen aanwezig zijn. Controleer ook grondig spelling en grammatica.

Respecteer de paginalimiet, bestandsnamen en deadline.

Referenties

[1] François Maniquet et al. *Wat heet dan gelukkig zijn? Geluk, welvaart en welzijn van de Belgen.* 2018.

Tabel 4: Getalwaarden rapporteren.

Statistiek	Vuistregel	Voorbeelden
Meting x_i	Volgens meetnauwkeurigheid.	$123\,\mathrm{mm}$, $123{,}4\,\mathrm{mm}$
Standaarddeviatie	Eén beduidend cijfer (meer in grote steekproeven).	$5\,\mathrm{mm}$
Schatting $\overline{x}, \hat{\beta}, \ldots$	Zelfde precisie als de standaardfout.	$9{,}812\,\mathrm{m/s^2}$, $5{,}97 \times 10^{24}\,\mathrm{kg}$
Standaardfout	Eén beduidend cijfer.	$0{,}001\,\mathrm{m/s^2}$, $0{,}05 \times 10^{24}\,\mathrm{kg}$
Percentage	In het algemeen geen decimalen,	$5\,\%$, $68\,\%$, $95\,\%$
	één beduidend cijfer voor waarde en complement.	$0{,}03\,\%$, $99{,}7\,\%$
t, F, χ^2, $\ldots$	Maximaal 1 decimaal en twee beduidende cijfers.	$t = -1{,}3$, $F = 11$, $\chi^2 = 4{,}1$
p-waarde	Eén beduidend cijfer, *nooit* nul,	$0{,}4$, $0{,}08$
	ongelijkheden voor grote of kleine waarden,	$> 0{,}9$, $< 0{,}001$
	wetenschappelijke notatie bij grote aantallen testen.	7×10^{-5}, 2×10^{-16}

Algemeen:

- Verzorg steeds de notatie van getallen, gebruik wetenschappelijke notatie waar nodig.
- Gebruik eenheden waar van toepassing.
- Afrondingen gelden louter voor rapportering, werk in berekeningen steeds met alle gevonden cijfers.
- Bovenstaande regels zijn richtinggevend, denk zelf na over de noodzaak van meer of minder cijfers.
- De rol van standaarddeviatie van een meting en standaardfout (standaarddeviatie van een statistiek, bijvoorbeeld het gemiddelde) is fundamenteel anders. De standaardfout geeft een idee over de nauwkeurigheid van de statistiek, terwijl de standaarddeviatie van een veranderlijke enkel een idee geeft over de spreiding van de verschillende waarden, niet over de precisie van één specifieke waarde. Vandaar de verschillende rol in het bepalen van het aantal beduidende cijfers.