

Undersøgelse af faktorer der påvirker test resultater

Frederik Rosendal Rytter

Aarhus Universitet

25. november 2021

Introduktion

I denne rapport analyseres et datasæt fra Kaggle indeholdende information om test resultater i tre forskellige tests, læsning, skrivning og matematik, og dertilhørende faktorer, såsom forældres uddannelses niveau, køn og information om hvorvidt prøvetageren deltog i test forberedelsen.

Specielt undersøges sammenhænge mellem disse faktorer, og herunder om der er en signifikant forskel i test resultater fra den generelle befolkning og prøvetagere, der deltog i test forberedelsen. Yderligere undersøges, om der er en signifikant forskel, i test resultater mellem prøvetagere, hvis forældre har forskellig uddannelsesbaggrund.

Rapporten er delt op i et data-afsnit, hvor dataet, samt indsamlingen af det gennemgås. Derefter et metode afsnit, hvor både teoretiske og praktiske redskaber brugt i analysen gennemgås, efterfulgt af analysen af dataet, som afrundes med en opsamling af de mest relevante resultater. Til sidst diskuteres de vigtigste overvejelser efterfulgt af en samlet konklusion på arbejdet.

1 Data

Data er indhentet fra Kaggle via følgende link: <https://www.kaggle.com/spscientist/students-performance-in-exams>. Data er blevet indsamlet af Jakki Seshapanpu. Datasættet består af otte søjler og tusind observationer, hvor søjlerne er givet som følgende:

- Køn
- Etnicitet
- Forældres uddannelses niveau

- Frokost
- Test forberedelses deltagelse
- Matematik test score
- Læse test score
- Skrive test score

Dataet er blevet rensat inden brug, hvor passende navngivninger af søjler er blevet foretaget. Direkte fra Kaggle, var dataet allerede næsten analyse klar. De første fem rækker af datasættet ser ud som følgende:

Unnamed: 0	gender	race/ethnicity	p_edu	lunch
0	female	group B	bachelor's degree	standard
1	female	group C	some college	standard
2	female	group B	master's degree	standard
3	male	group A	associate's degree	free/reduced
4	male	group C	some college	standard

testprep	math score	reading score	writing score
0	none	72	74
1	completed	69	88
2	none	90	93
3	none	47	44
4	none	76	78

2 Metoder

I analysen udregnes almene statistiske størrelser, herunder middelværdi og spredning. Viden om centralgrænseværdi sætningen bruges, til anvendelse af bootstrap-metoden til udledning af populations middelværdien - dette bruges i hypotese-testene i analysen. Et signifikans niveau på 0.01 antages på forhånd.

Til analysen bruges programmerings-sproget python og følgende biblioteker,

- Numpy
- Matplotlib
- Seaborn
- Scipy.stats

- Pandas

til udregning af numeriske størrelser, datahåndtering og visualisering.

3 Analyse

3.1 Testforberedelses deltagelse

Data'en antages som repræsentativ for den population der ligger til grund for den. Jvf. denne antagelse bruges bootstrap-metoden til approksimering af middelværdien af test-scorene for den population der ligger til grund for de observationer, der ikke deltog i testforberedelsen. Den udregnes til at være ca. 65.5 gennemsnitligt blandt de tre forskellige tests.

Middelværdien af test-scorene for den population der ligger til grund for de observationer, der deltog i testforberedelsen udregnes til at være ca. 74.4 gennemsnitligt blandt de tre forskellige tests.

Givet ovenstående antagelse opstilles følgende nul-hypotese:

"Middelværdien af test-scorene for samplen af observationer der deltog i testforberedelsen er ikke forskellig fra den af den generelle population."

og følgende alternative hypotese:

"Middelværdien af test-scorene for samplen af observationer der deltog i testforberedelsen er forskellig fra den af den generelle population."

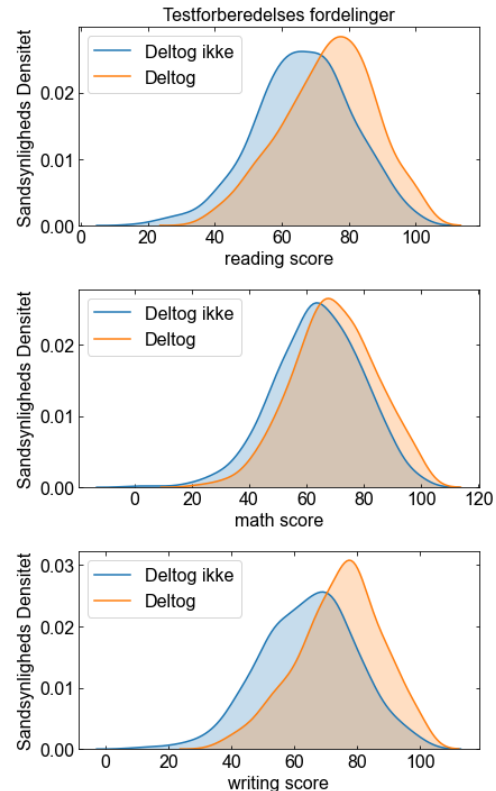
En visuel inspektion af de forskellige populations-fordelinger er givet i figur 1.

Givet den visuelle inspektion, forventes en signifikant forskel i middelværdi af henholdsvis læse-score og skrive-score, mellem populationen der deltog i testforberedelsen og populationen der ikke deltog i testforberedelsen.

Ved at udføre tre på hinanden følgende two-tailed one-sample t-test's og givet et signifikans niveau på 0.01, fås at forskellen mellem middelværdien af test-scorene for samplen af observationer der deltog i testforberedelsen, er signifikant forskellig fra den af den generelle population i både læse-, skrive- og matematik-testene.

3.2 Forældres uddannelses niveau

Data'en antages som repræsentativ for den population der ligger til grund for den. I disse udregninger bruges, for hver observation,



Figur 1: Kernel-Density-Estimering af sandsynlighedsfordelingen af score for populationerne der deltog i testforberedelsen 'Deltog' og dem der ikke gjorde 'Deltog ikke'.

middelværdien af de tre test-score. Jvf. denne antagelse bruges bootstrap-metoden til approksimering af middelværdien af test-scorene for den population der ligger til grund for de observationer, hvis forældres højeste uddannelses niveau er high-school. Den udregnes til at være ca. 63.1. Middelværdien af test-scorene for den population der ligger til grund for de observationer, hvis forældres højeste uddannelses niveau er en master, udregnes til at være ca. 73.6.

Givet ovenstående antagelse opstilles følgende nul-hypoteser:

(1) - *"Middelværdien af test-scorene for samplen af observationer, hvis forældres højeste uddannelses niveau er high-school, er ikke forskellig fra den af den generelle population."*

(2) - *"Middelværdien af test-scorene for samplen af observationer, hvis forældres højeste uddannelses niveau er en master, er ikke forskellig fra den af den generelle population."*

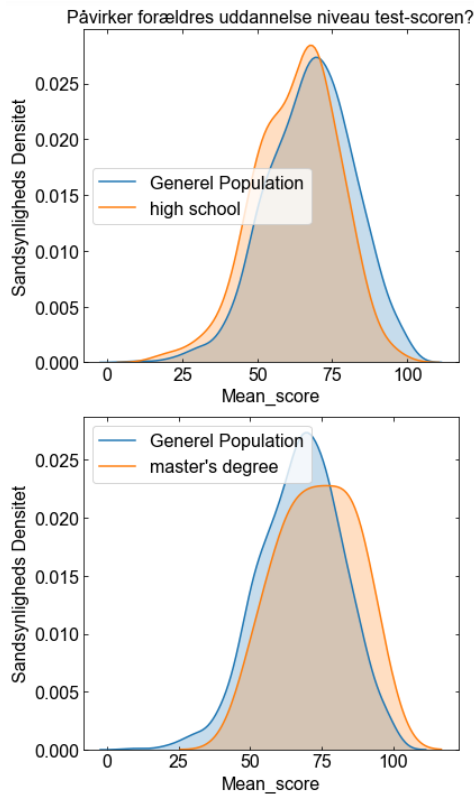
og følgende alternative hypoteser:

(1') - *"Middelværdien af test-scorene for samplen af observationer, hvis forældres højeste*

uddannelses niveau er high-school, er forskellig fra den af den generelle population."

(2') - "Middelværdien af test-scorene for samplen af observationer, hvis forældres højeste uddannelses niveau er en master, er forskellig fra den af den generelle population."

En visuel inspektion af de forskellige populations-fordelinger er givet i figur 2.



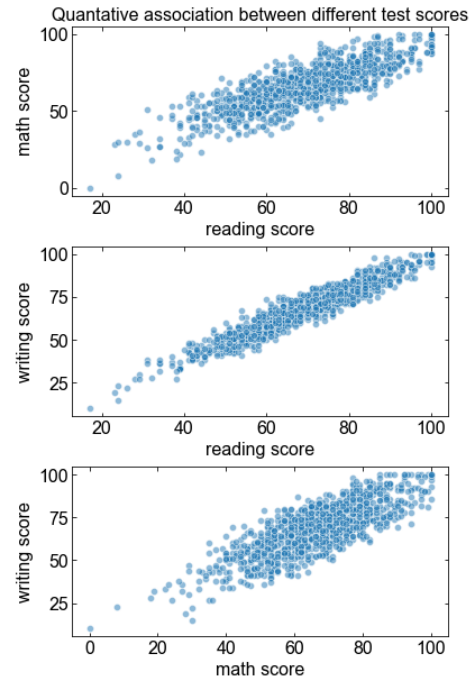
Figur 2: Kernel-Density-Estimering af sandsynlighedsfordelingen af score for populationerne, hvis forældres højeste uddannelses niveau er henholdsvis high-school og en master.

Givet den visuelle inspektion, forventes en signifikant forskel i middelværdi mellem populationen, hvis forældres højeste uddannelses niveau er en master, og den af den generelle population.

Ved at udføre to på hinanden følgende two-tailed one-sample t-test's og givet et signifikans niveau på 0.01, fås at forskellen mellem middelværdien af test-scorene for samplen af observationer, hvis forældres højeste uddannelses niveau er henholdsvis high-school eller en master, er signifikant forskellig fra den af den generelle population.

3.3 Kvantitativ associering mellem test-score

Til undersøgelse af hvorvidt der er en association mellem observationer, der klare sig godt i en test, og dermed også klare sig godt i de andre, plottes tre scatterplots, som kan ses i figur 3.



Figur 3: Scatterplot af den kvantitative associering mellem test-score.

Det er her oplagt, at der er en stærk lineær korrelation mellem observationer der klare sig godt i en test, og som dermed også klare sig godt i de andre tests.

4 Resultater

Et signifikans niveau på 0.01 blev valgt på forhånd, i det det forventes, at fem på hinanden følgende two-tailed one-sample t-test's skulle foretages, hvormed sandsynligheden for en type-1 fejl bliver,

$$1 - (1 - 0.01)^5 \approx 0.05 \quad (1)$$

svarende til en sandsynlighed på 5% for, at få et signifikant resultat, uden at det i virkeligheden er signifikant.

Givet hypotese-testene i analysen, kan det konkluderes at med 95% konfidens, at der er en signifikant forskel i middelværdien af test-scoren, for alle tre tests, mellem de observationer der deltog i test-forberedelsen og de

der ikke gjorde. Yderligere kan det med samme konfidens konkluderes, at der er en signifikant forskel i middelværdien af test-scoren, for alle tre tests, mellem de observationer, hvis forældres højeste uddannelses niveau er henholdsvis high-school eller en master, og den af den generelle population.

Til sidst blev det observeret, at der er en oplagt og stærk lineær korrelation mellem observationer der klare sig godt i en test, og som dermed også klare sig godt i de andre tests.

5 Diskussion

Vigtige faktorer at tage i betragtning er hvorvidt data'et var indsamlet tilfældigt, fra den overordnede population og om de individuelle observationer var uafhængige af hinanden.

Yderligere forventes det at data'et er normalfordelt og at det forekommer i en tilfredsstillende størrelse - som regel minimum 40 observationer. Alle disse faktorer påvirker troværdigheden af de endelige resultater og skal undersøges nøje.

Konklusion

I henhold til konklusionerne og argumenterne i afsnittet 'resultater', kunne det konkluderes at med 95% konfidens, at der er en signifikant forskel i:

- Middelværdien af test-scoren, for alle tre tests, mellem de observationer der deltog i test-forberedelsen og de der ikke gjorde.
- Middelværdien af test-scoren, for alle tre tests, mellem de observationer, hvis forældres højeste uddannelses niveau er henholdsvis high-school eller en master, og den af den generelle population.

Til sidst blev det i diskussionsafsnittet, diskuteret at vigtige faktorer at tage i betragtning er:

- Om data'et var indsamlet tilfældigt fra den overordnede population.
- Var de individuelle observationer var uafhængige af hinanden.
- Er data'et normalfordelt.
- Er der minimum 40 observationer - ellers skal andre statistiske redskaber bruges.

Alle disse faktorer påvirker troværdigheden af de endelige resultater og skal undersøges nøje.

Acknowledgements

Jakki Seshapanpu for indsamling af data. Link: <https://www.kaggle.com/spscientist/students-performance-in-exams>

Github

Link: <https://github.com/fred465f/exams>