DSAI_Samenvatting

May 29, 2022

1 Datascience & AI

Samenvatting voor examen van AJ 2021-2022. Door ydm#1001.

---

2 Module 1 Basisbegrippen, steekproefonderzoek

2.1 Basisbegrippen

Variabele = algemene eigenschap object, kan objecten onderscheiden. **Waarde** = Specifieke eigenschap, interpretatie van var.

Variabele	Waarde
Gender	Man
Hoogte	180 cm
Funny	Neen.

2.1.1 Meetniveaus

= variabele types. Bepalen beste analyse methode. (visualisatie, centrale tendens en spreiding, verband onderzoeken,...)

Kwalitatief = niet noodzakelijk numeriek. Beperkt aantal waardes.

Nominaal: categorieën zoals gender, ras, land, vorm,...
Ordinaal: Order, rank zoals militaire rank, onderwijsniveau,...

Kwantitatief = Numeriek met eenheid. Veel waardes die vaak uniek zijn.

Interval: Geen vast nulpunt => geen proporties. [1] (°C, °F)
Ratio: Absoluut nulpunt => wel proporties (bv afstand, energie, gewicht,...)

[1] 20 m is 1/3de (~33%) langer dan 15 meter (wel proportie) <-> 20°C is niet 1/3de warmer dan 15 °C (geen proportie)

Relaties tussen variabelen. variabelen hebben en verband als hun waardes systematisch veranderen.

	Pepsi	Coca Cola	Total
Like	56	24	80
Dislike	14	6	20
Total	70	30	100

Totalen zijn *Marginale totalen*

Onderzoek vaak naar **oorzakelijk verband** (frustratie lijd tot agressie, ...).
Oorzaak: onafhankelijke variabele
Verband: Afhankelijke variabele

Een verband tussen 2 variabelen zijn niet noodzakelijk oorzakelijk verband!

2.2 Steekproef

Populatie: Volledige verzameling objecten/personen die je wilt onderzoeken
Steekproef: Deel van de populatie waarop metingen uitgevoerd worden.
In bepaalde gevallen is het resultaat van de steekproef toepasbaar op de volledige populatie.

Steekproefmethode: Bepalen populatie -> bepalen steekproefgrootte -> Kiezen van steekproefmethode (budget en tijd)

Hoe keuze maken voor steekproef?
aselecte steekproef: elk element van de populatie heef evenveel kans om gekozen te worden.
Niet aselecte steekproef: De elementen van een sample zijn niet random gekozen. Objecten die makkelijker verkregen worden zijn waarschijnlijker om deel te nemen aan de steekproef. (convenience sampling genoemd).

Stratified to variables: populatie verdeeld op basis van een kenmerk (bijvoorbeeld leeftijd,...). (ook kan vgm bij dit voorbeeld alles /10 gedaan worden (zie slides voorbeeld) en is dit ook stratified)

Gender	<=18	]18,25]	]25,40]	>40	Totaal
Vrouw	500	1500	1000	250	3250
man	400	1200	800	160	2560
Totaal	900	2700	1800	410	5810

2.2.1 Fouten

	Steekproeffout	niet steekproeffout
Accidental	Puur toeval	Onjuist antwoord aangeduid

	Steekproeffout	niet steekproeffout
Systematisch	Online onderzoek: mensen zonder internet uitgesloten. Straat onderzoek: enkel die op dat moment daar aan het wandelen is Vrijwilligers onderzoek: enkel geïnteresseerde mensen	Slecht of niet gecalibreerd meetmateriaal Waarde beïnvloed door het feit dat je het meet. Antwoorders liegen (bv aantal sigaretten per dag)

2.3 Algemene imports.

[ ]:
```
#imports
import numpy as np                                   # "Scientific computing"
import scipy.stats as stats                          # Statistical tests

import pandas as pd                                  # Data Frame
from pandas.api.types import CategoricalDtype

import matplotlib.pyplot as plt                      # Basic visualisation
from statsmodels.graphics.mosaicplot import mosaic   # Mosaic diagram
import seaborn as sns                                # Advanced data␣
↪visualisation
import altair as alt                                 # Alternative visualisation␣
↪system
import math
from sklearn.linear_model import LinearRegression
```

2.4 Python Module 1

[ ]:
```
#Import data van een csv file
ais = pd.read_csv('../data/ais.csv')
#indien geen , maar bijvoorbeeld ; gebruikt dan is het
#pd.read_csv(fileLink, delimiter=';')

##Eerste aantal lijnen tonen
ais.head()

#Aantal rijen en kolommen in een dataset printen
print(f"Aantal rijen: {len(ais)}")
```

```
#Aantal kolommen
print(f"Aantal kolommen: {len(ais.columns)}")
#Algemene info over dataset.
ais.info()
#lijntje * printn
print("*"*50)
#Aantal kolummen per type
print(ais.dtypes.value_counts())

#kolom als index instellen
ais.set_index(['id'])

#Voor een kolom categorie als meetvariabele instellen
ais.sex = ais.sex.astype("category")

#Kan ook variabelen als ordinaal aanduiden met een ordening. Bijvoorbeeld als␣
↪we voor sex zouden doen.
# Voorbeeld:
print(ais.sex.unique()) #uniek
sex_Type = CategoricalDtype(categories=['f','m'], ordered=True) #en orderen
ais.sex= ais.sex.astype(sex_Type)
#een kolom beschrijven
print(ais.ferr.describe())

#SELECTEREN DATA
#Toon de tweede rij
ais.iloc[[1]]

#Toon rij 4 tot en met 6
ais.iloc[4:7]
#Toon KOLOM 6 tem 8: (ferr, bmi, ssf)
ais.iloc[:,5:8]
#Toon 1 variabelen (pcBfat)
ais['pcBfat']

#Toon alles van specifieke query (sport=netball)
ais.query("(sport=='Netball')")
#Toon specifieke colom met specifieke query (colom wt van sport=netball)
ais.query("(sport=='Netball')").wt

#Toon allesmet een bmi>26
print("BMI ding")
print(ais[ais.bmi>26])

#Toon frequentie en dergelijke
bmiais = ais[ais.bmi>26]
sns.countplot(x=bmiais.sport, data=bmiais)
```

```python
#Tel hoevaak een bepaalde categorie voorkomt
ais["Sport"].value_counts()
```

---

3 Module 2 Analyse van 1 variabele

3.1 Centrale tendens en spreiding

3.1.1 Maten van centrale tendens

Mean of Average De *arithmetic mean* is de som van alle waarden gedeeld door het aantal waarden. > $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Median Sorteer alle waarden en neem het middelste (gemiddelde bij een oneven).
Mode de mode is de waarde die het meest voorkomt in een dataset.

3.1.2 Maten van centrale spreiding

Range Absolute waarde van het verschil tussen het hoogste en laagste waarde.
Quartielen De quartielen van een gesorteerde set zijn 3 waarde die de set in 4 gelijke delen verdelen. Q_1, Q_2 , Q_3

Variantie: De variantie (S^2 of σ^2) is hetgemiddelde (mean) van het kwadraat van het verschil van de waardes van de dataset en het gemiddelde (arithmetic mean).
>$S^2 = \frac{1}{n-1}\sum_{i=i}^{n}(x_i - \overline{x})^2$

Standaard afwijking: De standaard afweiking (S of σ) is de wortel van de variantie

3.1.3 Samenvatting Centrale tendens en spreiding

Meetniveau	Center	Sprijdingsmaat
Kwalitatief	Mode	-
Kwantitatief	Avergae/mean Median	Variantie, standaard afwijking, range, interkwartielafstand

3.1.4 Samenvatting Symbolen

	Populatie	Steekproef
Aantal elementen	N	n
Gemiddelde (mean)	μ	$\overline{x}$
Variantie	$\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$	$S^2 = \frac{\sum(x_i-\overline{x})^2}{n-1}$

	Populatie	Steekproef
Standaard diviatie	σ	S

3.2 Data visualisatie

3.2.1 grafiek type overzicht

Meetniveau	Grafiek type
Kwalitatief	Staafdiagram
Kwantitatief	Boxplot Histogram Density plot

Taart diagrammen
vermijd gebruiken van taart diagrammen. Hoeken vergelijken is moeilijker dan lengtes, onbruikbaar voor veel categorieën

Tips Assen labelen, duidelijke titel, eenheid, label die de grafiek verduidelijkt.

Data distortion = zorgt voor fout interpreteren.

3.3 Python Module 2

[ ]:
```
#distributie van gevens  voor sport (distribution)
sns.displot(data= ais["sport"])

#categorie plot voor sports
sns.catplot(data= ais, kind="count", x="sport")

#distribution met Kernel density estimate (soort van normaalverdeling achtige␣
↪ding te krijgen)
sns.displot(data=ais[ais.sex=="f"].ht, kde=True)



#dingen
rowers = ais[ais.sport == "Row"].ht
print(f"Mean:                {rowers.mean()}")
print(f"Standard deviation:  {rowers.std()}") # Pay attention: n-1 in the␣
↪denominator
print(f"Variance:            {rowers.var()}") # Pay attention: n-1 in the␣
↪denominator
print(f"Skewness:            {rowers.skew()}")
print(f"Kurtosis:            {rowers.kurtosis()}")

# Median & co
```

```
print(f"Minimum:   {rowers.min()}")
print(f"Median:    {rowers.median()}")
print(f"Maximum:   {rowers.max()}")
percentiles = [0.0, 0.25, 0.5, 0.75, 1.0]
print("Percentiles", percentiles, "\n", rowers.quantile(percentiles))
print("Inter Quartile Range:", rowers.quantile(.75) - rowers.quantile(.25))
print(f"Range :    {rowers.max() - rowers.min()}")
```

---

4 Module 3.1 De centrale limietstelling, betrouwbaarheidsintervallen

---

4.1 Kansverdeling van een steekproef

4.1.1 Kans

Kans is de relatieve frequentie van het voorkomen van een bepaald event (bij uitvoeren van groot aantal onafhankelijke experimenten)
- kansen zijn getallen aan een set toegewezen - Die sets zijn deel van een allesomvattende set, het *universum* Ω - De nummers (kansen) toegewezen aan een set voeldoen aan 3 basis regels (axiom van kans) om overeen te komen met hoe kansen werken 1. Kansen zijn niet negatief $P(A) \geq 0 voorelkeA$. 2. Het universum heeft een kans 1: $P(\Omega) = 1$. 3. Als A en B disjunct zijn ($A \cap B = \varnothing$) dan geld $P(A \cup B) = P(A) + P(B)$ dit heet de somregel.

Eigenschappen 1. Complement regel: voor elke A geld $P(\overline{A}) = 1 - P(A)$ als $\overline{A}$ voorstelt dat A niet voorkomt. 2. Het onmogelijke event is kans nul: $P(\varnothing) = 0$ 3. De algemene som regel: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Onafhankelijke events Een event is onafhankelijk als het voorkomen van dit event (of het weten dat dit voorkomt) de kans dat een ander event gebeurt niet beïnvloed. Wiskundig: $P(A \cap B) = p(A)P(B)$

4.1.2 Random variabele

Een random variable is een waarde toekennen aan verschillende gebeurtenissen. Bijvoorbeeld. 1 als je een J trekt uit een kaart spel, 2 bij een Q en 3 bij een K en tot slot 0 bij alle andere mogelijkheiden Kansverdeling functie (PDF) wiskundig. $f_x(x) = P(X = x)$
Voorbeeld:

Expectation of a R.V. Verwachting van een random variabele is geschreven door μ_x of $E(X)$ en is gegeven door $\mu_X = \sum_i x_i P(X = x_i) = \sum_i x_i f_x(x_i)$

Variantie van een R.V. De variantie van een random variabele is bepaald door $\sigma^2 = \sum_i (x_i - \mu_x)^2 P(X = x_i) = \sum_i (x_i - \mu_x)^2 f_x(X_i)$.
Standaard afwijking: $\sigma_x = \sqrt{\sigma_x^2}$

Continue random variabele

- Een continue R.V. neemt een ontelbaar oneindig aantal mogelijke waardes
- In dat geval niet logisch dat de kans van X = a exact te bekijken, omdat de kans altijd 0 is.
- Wat wel zin heeft is te bekijken wat de kans is van X=[a,b].
- Deze kans kan gevonden worden door te integrating de PDF van random variabale.
- voorbeeld van Continue rando value: lengte van mensen in een populatie...

De lengte van mensen volgt vaak ongeveer een **Normale verdeling**. De normale verdeling is een type van **Continue kansverdeling**. De formules voor de variantie en de verwachting zijn dezelfde als voor gewone R.V.'s maar dan met een integraal van -oneindig tot +oneindig.

4.1.3 Standaard normaal verdeling

x en z hebben gelijkaardige positie in de gauss curve. Wat is de wiskundige relatie tussen x en z?

$x = \mu + z * \sigma$ and $z = \frac{x-\mu}{\sigma}$

- $[-1, 1]$ ($[\mu - \sigma, \mu + \sigma]$) bevat 68,3% van de populatie of kans
- $[-2, 2]$ ($[\mu - 2\sigma, \mu + 2\sigma]$)bevat 95,4%
- $[-3, 3]$ ($[\mu - 3\sigma, \mu + 3\sigma]$) bevat 99,7%

4.1.4 Exponentiele spreiding

is een andere veelgebruikte continue distributie. Dit gebeurt als er minder grote waarden zijn en meer kleine waarde. Bijvoorbeeld het bedrag dat klanten uitgeven volgt een exponentiele distributie. Er zijn meer mensen die kleine bedragen uitgeven dan mensen die veel uitgeven.

4.1.5 Continue uniforme spreiding

beschrijft een experiment waar een arbitreire uitkomst is tussen bepaalde grenzen. de density functie is constant omdat er voor elke waarde een even grote kans is dat ze voorkomt. Bijvoorbeeld een lift gaat altijd tussen de 10 en 15 seconden naar de 2de verdieping dan is de kans altijd 1 dat je binnen de 10 en 15 seconden op dat verdiep bent. (neem de trap is beter voor je gezondheid!).

4.2 Van steekproef naar populatie

4.2.1 De centrale limietstelling

Als de steekproefgrootte groot genoeg is dan zal de kansverdeling van het steekproefgemiddelde ongeveer een normale verdeling zijn, onafhankelijk van de kansverdeling van de onderliggende populatie.
Bekijk een random steekproef met $ n$ observaties uit een populatie met verwachte waarde μ en standaard deviatie σ. Als $ n$ groot genoeg is dan zal de kansdichtheid van de steekproefgemiddelde $\overline{x}$ ongeveer een normale verdeling met gemiddelde $\mu_{\overline{x}}$ en een standaardafwijking $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$. Hoe groter de steekproef hoe beter de kansverdeling van $\overline{x}$ zal benaderen met verwachte waarde van de populatie, μ.

4.2.2 Punt schatting

Een punt schatting van een populatie parameter is een formule of vergelijking die toestaat om een verwachte waarde te bepalen voor die parameter.

4.2.3 Confidence interval

Een confidence interval is een vergelijking of formule die toestaat een interval op te stellen die met een bepaalde zekerheid een parameter bevat.

Voor kleine steekproef is de centrale limietstelling niet geldig. In de plaats daarvan zeggen we als de populatie X een normale verdeling heeft en je hebt een kleine steekproef met $\overline{x}$ en standaard afwijking s dan $t = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$ zal gedragen als een t-distributie met $ n-1 $ graden van vrijheid.

4.3 Python Module 3.1

[ ]:
```
#voorbeeld incl print met 3 cijfers na komma.

#kansberekening voor normale verdeling met
m = 0 #gemiddelde
s = 1 #standaard afwijking
print('P(Z<1.33)=%.3f'%stats.norm.cdf(1.33, loc=m, scale=s))
print('P(Z>1.33)=%.3f'% stats.norm.sf(1.33, loc=m,scale=s))
print('P(-1.35<Z<-0.10)=%.3f'% (stats.norm.cdf(-0.10, loc=m,scale=s)-stats.norm.
 ↪cdf(-1.35, loc=m,scale=s)))


## Probability density function (blauw) ingekleur
m=2.5
s= 1.5
dist_x = np.linspace(m - 4 * s, m + 4 * s, num=201)
dist_y = stats.norm.pdf(dist_x, m, s)
plt.plot(dist_x, dist_y)
plt.fill_between(dist_x, 0, dist_y, color='lightblue')
#zelfde oef de cdf (oranje)
dist_y_cdf = stats.norm.cdf(dist_x)
plt.plot(dist_x, dist_y_cdf)
#de area onder de pdf tussen 0.5 en 4
stats.norm.cdf(4, loc=m,scale=s)-stats.norm.cdf(0.5, loc=m,scale=s)


#Genereer random nummers die de standaard normaal verdeling volgen
n=25
observations = np.random.normal(loc=m, scale=s, size=n)
#print een histogram met kansdichtheids functie en theoretishe kansdichtheids
sns.histplot(observations, kde=True)

#Bij standaardafwijking en gemiddelde van populatie naar s en m van een␣
 ↪steekprof te gaan (n= steekproefgrootte)
n=81
```

```
standaarddivPop= 36
s= standaarddivPop/math.sqrt(n) #s= standaarddiv van steekproef voor dat␣
 ↪gemiddelde
```

---

5 Module 3.2 Hypothesetesten

5.1 Testprocedure

5.1.1 Statistische hypothesetesten

- **Hypothesis**: idee dat nog moet bewezen worden: statement over een numerische waarde van een populatie parameter.
- **Hypothesetest**: Verificatie van het statement over de waardes van 1 of meerdere populatie parameters.
- **Null Hypothese (H_0)**: Basis hypothese, aan nemen dat die waar is
- **Alternatieve Hypothese (H_1, H_a)**: Conclusie als de null hypothese waarschijnlijk fout is.

5.1.2 Elementen van een test procedure

- **Test statistiek**: De waarde die berekend word van een steekproef
- **Acceptatieregio**: De regio van waardes die de null hypothesis bevestigen
- **Kritieke regio/ Regio van afwijzing**: De regio van waardes die de null hypothesis verwerpen.
- **Significantie niveau**: De waarschijnlijkheid van verwerpen van de null hypothese H_0

5.1.3 Test procedure

1. Formuleer beide hyptoheses (H_0, H_1)
2. Bepaal significantie niveau (α)
3. Berkenen test statistiek
4. Bepalen kritieke regio van de kans waarde
5. Conclusies trekken.

5.2 Kans waarde (Probability value)

p-waarde: De p-waarde is de kans, als de null hyptohese waar is, een waarde voor de test statistiek te krijgen die minimaal zo extreem is als de geobserveerde waarde.

- p-waarde $< \alpha$ => verwerpen H_0: de ontdekte waarde voor $\overline{x}$ is te extreem.
- p-waarde $\leq \alpha$ => H_0 niet verwerpen: De ontdekte waarde voor $\overline{x}$ kan nogsteeds uitgelegd worden door toeval.

5.3 Kritieke regio

De Kritieke regio is de verzameling van alle waarden van een test statistiek waarvoor de null hypothese verworpen kan worden.

Kijk naar kritieke waarde g waarvoor geld: P(M>g)= α
Bepaal z_α waarvoor geldt: $P(Z > z_\alpha) = \alpha => g = \mu + z_\alpha * \frac{\sigma}{\sqrt{n}}$

- Links van G: Regio van acceptatie (H_0 niet verwerpen)
- Rechts van G: Kritieke regio (H_0 verwerpen)

5.3.1 Samengevat Testing procedures

Goal	Test met betrekking tot de waarde van de populatie gemiddelde μ gebruik makend van steekproef van n onafhankelijke waardes.
Voorwaarde	De populatie heeft een random verdeling, n is groot genoeg

Test type	Two-tailed	Left-tailed		Right-tailed
H_0	$\mu = \mu_0$	$\mu = \mu_0$		$\mu = \mu_0$
H_1	$\mu != \mu_0$	$\mu < \mu_0$		$\mu > \mu_0$
Critieke regio	$\lvert\overline{x}\rvert > g$	$\overline{x} < -g$		$\overline{x} > g$
Test statistiek			$z = \frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	

5.3.2 Voorwaardes voor z-test

- De steekproef moet aselect zijn
- De steekproefgrootte moet groot genoeg zijn ($n \geq 30$)
- De test statistiek moet zich gedragen als een normale verdeling
- De tandaard diviatie van de populatie σ is gekend.

Soms zijn de voorwaardes niet voldaan en dan kan geen z-test

5.4 Voorbeelden

zie slides.

5.5 Student's t-test

Wat als voorwaardes voor z-test niet voldaan zijn? - steekproef niet groot genoeg - populatie standaard diviatie niet gekend Als de variabele normaal verdeeld zijn kunnen we de t-test gebruiken.

5.5.1 De t-test

Kritieke waarde bepalen:
$g = \mu \pm t * \frac{s}{\sqrt{n}}$

- t-waarde afgelijd uit de Student t-distributie gebaseerd op n-1 vrijheidsgraden
- De waarde opzoeken door `t.isf` in Python
- Los van dit is de procedure gelijk aan de procedure voor de z-test.

5.6 Fouten in hypothese testen

	Rea	liteit
Conclusion	H_0 True	H_1 True
H_0 niet verworpen	correct conclusie	Type II fout (vals negatief)
H_0 verworpen	Type I fout (vals positief)	correct conclusie

P(type I error) = α (=significantie level)
P(type II error) = β
β berekenen is niet triviaal maar als α afneemt dan neemt β toe.

5.7 Python Module 3.2

[ ]:
```
#voorbeeld met Z
m = 44 #gemiddelde populatie
s = 6.2 #standaard div populatie
n = 72 #steekproefgroote
m_intro = 46.2 #steekproef gemiddelde
s_intro = s/math.sqrt(n) #steekproef std?
a = 0.025 #alpha waarde, significantie niveau

#Plot de grafiek voor ja de normaal verdeling enz
dist_x = np.linspace(m-4*s_intro,m+4*s_intro, num=201)
dist_y = stats.norm.pdf(dist_x, m, s_intro)
plt.plot(dist_x, dist_y)

#Inkleuren het stuk dat beter is? ofja gwn het te onderzoeken het H0 ding
plt.fill_between(dist_x, 0, dist_y, where=(dist_x>=m_intro), color='red')
#vertikale lijn daarvoor
plt.axvline(m_intro, color="green")

#p waarde berekenn
p_waarde = stats.norm.sf(m_intro, loc=m, scale=s_intro)
print("p waarde: %.4f"%p_waarde)
if(p_waarde < a):
    print("p < a: reject H0")
else:
    print("p > a: do not reject H0")
#kritieke regio bepalen
g_value = stats.norm.isf(0.025, loc = m, scale= s_intro)
print("Critical value g   %.3f" % g_value)
if (m_intro < g_value):
    print("sample mean = %.3f < g = %.3f: do not reject H0" % (m_intro,␣
 ↪g_value))
else:
```

```
    print("sample mean = %.3f > g = %.3f: reject H0" % (m_intro, g_value))

#voorbeeld met t
#H0: dat het niet significant groter is
prijs= [400, 350, 400, 500, 300, 350, 200,
 500, 200, 250, 250, 500, 350, 100] #is gwn dataset steekproef
m = 300 #mag ni significant groetr zijn dan
a=0.05 #significantieniveau
n = len(prijs)  #lengte sample
m_samp = np.mean(prijs) #gemiddelde samp
s= np.std(prijs, ddof=1) #std samp
sn = s/math.sqrt(n) #s/wortel(n)

print(f"mean is {m_samp}")
print(f"std is {s}")

#tekenen
dist_x = np.linspace(m-4*sn,m+4*sn, num=201)
dist_y = stats.t.pdf(dist_x, loc=m, scale=sn, df=n-1)
plt.plot(dist_x, dist_y)
#aanduiden welk deel "oke" is voor H0
plt.fill_between(dist_x, 0, dist_y, where=(dist_x<=m_samp), color='yellow')
plt.axvline(m_samp, color="blue")

#p waarde
p = stats.t.sf(m_samp, loc=m, scale = sn, df = n-1)
print("p:value %.5f"% p)
if(p<a):
  print("p<a:rejectH0")
else:
  print("p>a: niet reject H0")

#kritieke waarde (g)
g = stats.t.isf(a, loc=m, scale=sn, df=n-1)
print('critiek g ong= %.3f' % g)
if (m_samp < g):
    print("sample mean = %.3f < g = %.3f: do not reject H0" % (m_samp, g))
else:
    print("sample mean = %.3f > g = %.3f: reject H0" % (m_samp, g))
```

---

6 Module 4 Analyse van 2 kwalitatieve variabelen

6.1 Bivariate Analyse

- Is bepalen of er een verband is tussen 2 stochastische variabelen (X en Y)

- **Verband** = je kan voorspellen (tot zekere hoogte) wat de waarde van Y is op basis van de waarde van X
  - X is onafhankelijke variabele
  - Y is afhankelijke variabele
- **Opgelet** Een verband is niet noodzakelijk een oorzakelijk verband.

6.1.1 Overzicht bivariate analyse

Onafhankelijke	**Afhankelijke**	**Test/Metric**
Kwalitatief	Kwalitatief	χ^2-test Cramér's V
Kwalitatief	Kwantitatief	Two-sample t-test Cohen's d
Kwantitatief	Kwantitatief	- Regression, correlation

6.2 Contingency tables

(= tabel waarin ene variabele in de rijen en een andere in de kolom om zo een verband te onderzoeken.)

Gender Survey	Female	Male	Total
Strongly disagree	0	4	4
Disagree	17	45	62
Neutral	23	91	114
Agree	12	53	65
Strongly agree	0	5	5
Total	52	198	250

6.2.1 Verwachte waardes

Als er geen verschil (associatie) is verwachten we dezelfde ratios in elke kolom van de tabel. In elke cel dus: (rij totaal x kolom totaal)/n

Gender Survey	Female	Male	Total	EFemale	EMale	ETotal
Strongly disagree	0	4	4	0.832	3.168	4
Disagree	17	45	62	12.896	49.104	62
Neutral	23	91	114	23.712	90.288	114
Agree	12	53	65	13.520	51.480	65
Strongly agree	0	5	5	1.040	3.960	5
Total	52	198	250	52	198	250

voorbeeld Strongly Disagree:
EFemale = $\frac{52(FemaleTot)*4(Stronglydisagreetot)}{250(totaaltotaal)} = 0.832$

EMale $=\frac{198(MaleTot)*4(Stronglydisagreetot)}{250(totaaltotaal)} = 3.168$
Totaal Strongly disagree = 0.832 + 3.168 = 4 (= strongly disagree tot)

6.2.2 Sprijding meten

Hoe ver is de geobserveerde waarde van de verwachte e? $\frac{(o-e)^2}{e}$.

6.2.3 De chi-kwadraad statistiek

$\chi^2 = \sum_i \frac{(o_i-e_i)^2}{e_i}$

- o_i = aantal observaties in de i'de cel van de contingency tabel
- e_i = verwachte frequentie
- Kleine waarde => geen verband
- Grote waarde => verband

Wanneer χ^2 groot genoeg? - bij 2x2 tabel is $\chi^2 = 10$ relatief groot => verband - bij 5x5 tabel is $\chi^2 = 10$ relatief klein => geen Verband

Er is dus nood aan een ding onafhankelijk aan de groote van de tabel

6.2.4 Cramér's V

$V = \sqrt{\frac{\chi^2}{n(k-1)}}$
Met n het aantal observaties en k het min(numRows, numCols)

Cramér's V	Interpretatie
≈ 0	Geen verband
≈ 0.1	Zwak verband
≈ 0.25	Gemiddeld verband
≈ 0.5	Sterk verband
≈ 0.75	Heel sterk verband
≈ 1	Volledig verband

6.3 Chi-kwadraad test voor onafhankelijkheid

- = alternateif voor Cramér's V om te onderzoeken wat het verband is tss kwalitatieve variabelen
- De waarde van χ^2 verdeeld over de χ^2 verdeling?

6.3.1 Test procedure

1. Hypotheses opstellen:
   - H_0: er is geen verband (χ^2 is klein)
   - H_1: er is een verband (χ^2 is groot)
2. Kies significantie niveau
3. Bereken teststatistiek (χ^2)
4. Gebruik df = (numRox-1)*(numCol-1) en ofwel:
   - Kritieke waarde g berkenen (P(χ^2)=α)

- Bereken p-waarde

5. Trek conclusies
    - $\chi^2 < g$: H_0 niet verwerpen; $\chi^2 > g$: H_0 verwerpen
    - p $> \alpha$: H_0 niet verwerpen; p $< \alpha$: H_0 verwerpen

In python:

[ ]:
```
def chiAndPFromContingency(data):
  observed=pd.crosstab(data.colmn1, data.colmn2)
  chi2, p, df, expected = stats.chi2_contingency(observed)
  return chi2,p,df, expected
```

6.4 Goodness-of-fit test

De χ^2 test kan ook gebruikt worden voor bepalen van representatieviteit van een steekproef voor de populatie.
De goodness-of-fit test indiceert in welke mate een steekproef overeenkomt met de null hypothese met betrekking tot de verdeling van een kwalitatieve var over meerdere exclusieve klassen.
Als χ^2 klein is dan is het representatief en anders niet. χ^2 meet tot welke mate er conflict is met de null hypothese

6.5 Gestandardiseerde residuen

Indiceert welke klassen de meeste bijdragen bieden tot de waarde van χ^2

$r_i = \frac{o_i - n\pi_i}{\sqrt{n\pi_i(1-\pi_i)}}$

- $r_i \in [-2, 2]$ Acceptabel
- $r_i <$ -2 = onder vertegenwordigd
- $r_i >$ 2 = over vertegenwoordigd

6.6 Cochran's rules

Om χ^2 test te kunnen toepassen moeten de volgende regels voldaan zijn: 1. Van alle categorieen moet de verwachte frequentie e hoger zijn dan 1 2. In maximum 20% van de categorieen is de verwachte frequentie e kleiner dan 5

6.7 Python Module 4

[ ]:
```

```

7 Module 5: Analyse van Kwalitatieve vs kwantitatieve variabelen

7.1 Data visualisatie

Chart types voor kwantitatieve data gegroepeerd per kwalitatieve data: Grouped boxplot, grouped density plot, bar chart MET ERROR BARS,…

7.2 Two sample t-test

Is het gemiddelde van 2 steekproeven significant verschillend?

7.2.1 Onafhankelijke steekproeven

met `stats.ttest_ind(a, b, alternative="less", equel_var=False)`
idk
Voorbeeld: Een clinische studie waarbij een groep een placebo krijgt en eeen andere groep het medicijn. #### Test procedure 1. Hypotheses (in voorbeeld): 1. $H_0 : \mu_2 - \mu_2 = 0$ 2. $H_0 : \mu_2 - \mu_2 < 0$ 2. Significantie niveau: $\alpha = 0.05$ 3. Test statistiek: 1. $\overline{x} - \overline{y} = -12.833$ 2. $\overline{x} = estimation for \mu_1$ (control groep) 3. $\overline{y} = estimation for \mu_2$ (intervention groep) 4. p Berkenen 5. conclusie

7.3 Afhankelijke steekproeven

Bijvoorbeeld. Brandstof met toevoegingen lager verbruik of niet? 10 autos zonder toevoegingen en 10 met. `stats.ttest_rel(regular, additives, alternative='less')` #### Test procedure 1. Hypotheses (voorbeeld): 1. $H_0 : \overline{x - y} = 0$ 2. $H_0 : \overline{x - y} > 0$ 2. Significantie niveau: $\alpha = 0.05$ 3. Test statistiek: 1. $\overline{x - y}$ 2. x = mijl per lieter met additieve 3. y = mijl per lieter met gewoon 4. berkenen p 5. conclusie

7.4 Effect groote

De effectgrootte is een maat die uitdrukt hoe groot het verschil is tussen twee groepen. - control groep vs interventie groep - Kan gebruikt worden bovenop hypothese test - vaak gebruikt in onderwijs wetenschap - Er zijn meerdere definities. Hier: Cohen's d

7.4.1 cohen's d

d=$\frac{\overline{x_1} - \overline{x_2}}{s}$

met steekproefgemiddelde $\overline{x_1}, \overline{x_2}$ en s als standaard diviatie van beide groepen gecombineerd.
$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ met $ n_{1}, n_{2} $ de steekproefgrootte en s_1, s_2 de standaard afwijking

Interpretatie van cohen's d

$\lvert d \rvert$	Effectgrootte
0.01	Heel klein
0.2	Klein
0.5	Gemiddeld
0.8	Groot
1.2	Heel groot
2.0	Immens

in educational science: 0.4 = tipping punt voor gewenst effect. d= 1: verwerken leermateriaal van 1y in 6 maand. ### Typische aanpak in onderwijs onderzoek - Onderzoekvraag: is X een goede leermethode, heeft het dus positief effect - control groep gebruikt "gebruikelijke" methode - interventie groep gebruikt x - evaluatiemomenten - punten bepalen en d berekenen

---

7.5 Python Module 5

```
a,b = ["data"]
# Onafhankelijke steekproeven
stats.ttest_ind(a=a, b=b, alternative="less", equel_var=False)
### Afhankelijke steekproeven
stats.ttest_rel(a, b, alternative='less')
```

8 Module 6: Analyse van 2 kwantitatieve variabelen

8.1 Data visualization

Scatterplot - x-ass: onafhankelijke variabele - y-ass: afhankelijke variabele - elk punt is een observatie…

8.2 Lineaire regressie

8.2.1 Regressie

Met regressie proberen we een consistent en systematisch verband te vinden tussen 2 kwantitatieve variabelen: 1. **Monotoon**: consistente richting van de relatie tussen 2 variabelen: toenemen/afnemen 2. **Niet-monotoon**: Waarde van afhankelijke var veranderd systematisch met de waaarde van de onafhankelijke variabele maar de richting is niet consistent

8.2.2 Lineaire regressie

Characteristieken: - Aanwezigheid: is er een relatie? - Richting: toenemen of afnemen? - Sterkte van de relatie: sterk, gemiddeld, ZWAK, niet bestaande

8.2.3 Method of least squares

De regressielijn heeft de volgende vergelijking: $\hat{y} = \beta_1 x + \beta_0$ met:
- $\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$ - $\beta_0 = \overline{y} - \beta_1\overline{x}$

8.3 Covariantie

is de maat die indiceert of een verband tussen twee variabelen toeneemt of afneemt.
Cov(X,Y) $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$
- Cov > 0: neemt toe - Cov ≈ 0: geen relatie - Cov < 0: neemt af

Opmerking: Covariantie van populatie (met n in de noemer) vs van steekproef (met n-1 in de noemer)

8.4 Person’s correlatie coefficient

Person’s prduct-mooment correlatie coefficient R is een maat om de sterkte van een lineaire correlatie tussen x en y te meten.

R = $\frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$

R $\in$ [-1, +1]

8.5 Coefficient van determinatie

R^2 verklaart de percentages van de variantie van de geobserveerde waardes relatief tov de regressielijn.
R^2 = percentage variantie observaties verklaard door regressie lijn
1-R^2 = Percentage variantie observaties niet uitgelegd door de regressie

$\|R\|$	R^2	Explained variance	interpretatie
0.3	0.1	< 10%	Heel zwak
0.3 - 0.5	0.1 - 0.25	10 - 25%	Zwak
0.5 - 0.7	0.25 - 0.5	25 - 50%	Gemiddeld
0.7 - 0.85	0.5 - 0.75	50 - 75%	Sterk
0.85 - 0.95	0.75 - 0.9	75 - 90%	Heel sterk
> 0.95	> 0.9	> 90%	Uitzonderlijk

8.6 Bedenkingen

- De correlatie coefficient kijkt enkel naar het verband tussen 2 variabelen. Interacties met andere variabelen wordd niet in rekening genomen.
- De correlatie coefficient gaat expliciet NIET uit van een oorzakelijk verband
- Pearson's correlatie coefficient toont enkel lineaire verbanden

---

8.7 Python Module 6

[ ]:
```
cats= ["een dataset met onafh var Hwt en afh var Bwt"]
#scatterplot
sns.relplot(data=cats,x='Hwt', y='Bwt', hue=cats.Sex)
sns.scatterplot(data=cats,x='Hwt', y='Bwt', hue=cats.Sex)
#scatterplot met regressie lijn
sns.regplot(x=cats.Hwt, y=cats.Bwt)
xwaarde = cats.Hwt.values.reshape(-1,1)
ywaarde = cats.Bwt
rets = LinearRegression().fit(xwaarde, ywaarde)
print(f"Regressielijn: y^={rets.intercept_:.2f} + {rets.coef_[0]:.2f}x")

#correlatie coefficient
cor = np.corrcoef(cats.Hwt, cats.Bwt)[0][1]
# Determination
det = cor**2
```

---

9 Samenvatting van enkele zaken

9.1 Samenvatting Symbolen

	Populatie	Steekproef
Aantal elementen	N	n
Gemiddelde (mean)	μ	$\overline{x}$
Variantie	$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$
Standaard diviatie	σ	S

	Symbool	Formule
Expectation van een random value	μ_x of $E(X)$	$\mu_X = \sum_i x_i P(X = x_i) = \sum_i x_i f_x(x_i)$
Variantie van een random value	σ^2	$\sum_i (x_i - \mu_x)^2 P(X = x_i) = \sum_i (x_i - \mu_x)^2 f_x(X_i)$
Standaardafwijking van een R.V.	σ_x	$\sqrt{\sigma_x^2}$

	Symbool	Tailed	Formule
		Rechts	$g = \mu + z_\alpha * \frac{\sigma}{\sqrt{n}}$
Kritieke regio	g	Links	$g = \mu - z_\alpha * \frac{\sigma}{\sqrt{n}}$
		2 zijdig:	$g = \mu \pm z_\alpha * \frac{\sigma}{\sqrt{n}}$

9.2 Overzicht bivariate analyse

Onafhankelijke	**Afhankelijke**	**Test/Metric**
Kwalitatief	Kwalitatief	χ^2-test Cramér's V
Kwalitatief	Kwantitatief	Two-sample t-test Cohen's d
Kwantitatief	Kwantitatief	- Regression, correlation

9.2.1 Cramér's V

$V = \sqrt{\frac{\chi^2}{n(k-1)}}$

Met n het aantal observaties en k het min(numRows, numCols)

En $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$

Cramér's V	Interpretatie
≈ 0	Geen verband
≈ 0.1	Zwak verband
≈ 0.25	Gemiddeld verband

Cramér's V	Interpretatie
≈ 0.5	Sterk verband
≈ 0.75	Heel sterk verband
≈ 1	Volledig verband

9.3 Samenvatting python code (zoals in slides in overzicht staan)

9.3.1 Normaalverdeling met mean m en standaard deviation s

Function `stats.`	Doel
`norm.pdf(x, loc=m, scale=s)`	kansdichtheid bij X
`norm.cdf(x, loc=m, scale=s`	Links kans (left-tail) $ P(X<x)$
`norm.sf(x, loc=m, scale=s`	Rechts kans (right-tail) $ P(X>x)$
`norm.isf(1-p, loc=m, scale=s)`	p% van observaties die lager verwacht werden dan het resultaat

9.3.2 Student t-verdeling

(df= degres of freedom) |Function `stats.` | Betekenis | |——|——| |`t.pdf(x, df=d)`| kansdichtheid voor x| |`t.cdf(x, df=d)`| Left-tail kans $ P(X<x)$| |`t.sf(x, df=d)`| Right-tail kans $ P(X>x)$| |`t.isf(1-p,df=d)`| p% van observaties is de verwachting lager als het resultaat|

9.3.3 χ^2 verdeling in python

Function `scipy.`	Betekenis
`chi2.pdf(x, df=d)`	kansdichtheid voor x
`chi2.cdf(x, df=d)`	Left-tail kans $ P(X<x)$
`chi2.sf(x, df=d)`	Right-tail kans $ P(X>x)$
`chi2.isf(1-p,df=d)`	p% van observaties is de verwachting lager als het resultaat

9.3.4 two sample t-test

Onafhankelijke steekproeven

[ ]:
```
stats.ttest_ind(a, b, alternative="less", equel_var=False)
```

9.3.5 Afhankelijke steekproeven

[ ]:
```
stats.ttest_rel(regular, additives, alternative='less')
```

9.4 Enkele functies

[ ]:
```
def kansdichtheidNorm(x, mean, standardDiviatie):
  return stats.norm.pdf(x, loc=mean, scale=standardDiviatie)
```

```python
def leftTailNorm(x, mean, standardDiviatie):
  return stats.norm.cdf(x, loc=mean, scale=standardDiviatie)

def rightTailNorm(x, mean, standardDiviatie):
  return stats.norm.sf(x, loc=mean, scale=standardDiviatie)

def zScore(x, mean, standardDiviatie):
  return (x-mean)/standardDiviatie

#Confidence interval Large sample (met z)
def confIntervalLarge(confidenceLevel, meanSample, standaardDivSample):
  alpha = 1-confidenceLevel
  p = 1-alpha/2
  zAlpha2Kans = stats.norm.isf(1-p)
  return [meanSample-zAlpha2Kans*(standaardDivSample), meanSample+zAlpha2Kans*(standaardDivSample)]

#confidence interval small sample (met t)
def confIntervalSmall(confidenceLevel, meanSample, standaardDivSample, sampleSize):
  alpha = 1-confidenceLevel
  p = 1-alpha/2
  zAlpha2Kans = stats.t.isf(1-p, df=(sampleSize-1))
  return [meanSample-zAlpha2Kans*(standaardDivSample/math.sqrt(sampleSize)), meanSample+zAlpha2Kans*(standaardDivSample/math.sqrt(sampleSize))]


# Chi^2 en p warade en dergeljike van een data op basis van colom
def chiAndPFromContingency(colom1, colom2):  #bv rlanders.Survey, rlanders.Gender
  observed=pd.crosstab(colom1, colom2)
  chi2, p, df, expected = stats.chi2_contingency(observed)
  return chi2,p,df, expected
```