Statische analyse van Bevolkingsdata (Gemeente Utrecht)

Applied Artificial Intelligence

Analytical Computing

Berry Hijwegen

Inhoud

Gemiddelde

Om het gemiddelde te berekenen wordt de som van alle elementen in de lijst gedeeld door de het aantal elementen in de lijst.

```
def length(lst):
    l = 0
    for item in lst:
        l += 1
    return l

def total(lst):
    _sum = 0
    for item in lst:
        _sum += item
    return _sum

def average(lst):
    return total(lst) / length(lst)
```

Mediaan

De mediaan is de middelste waarde. De lijst wordt hiervoor eerst gesorteerd.

Bij een oneven getal is dit vrij simpel, hierbij wordt de lengte van de lijst gedeeld door 2. Dit is de index van het middelste element.

Bij een lijst met een even aantal getallen is dit iets lastiger, hierbij moet het gemiddelde genomen worden van de twee middelste getalen. Daarom wordt hiervoor de lengte gedeeld door 2 – 0.5 en + 0.5. Vervolgens wordt van deze getallen op de twee indexen het gemiddelde berekend.

```
def median(lst):
    l = length(lst)
    if l % 2 == 0:
        total = (lst[int((l) / 2 - 0.5)]) + (lst[int((l) / 2 + 0.5)])
        return total / 2
    elif l % 2 == 1:
        return lst[int((l)/2)]
```

Modus

De modus is de meest voorkomende waarden. Om de modus te berekenen wordt er voor elke waarde een counter aangemaakt, de hoogste wordt gereturned.

```
def modus(lst):
    hits = {}
    highest = []
    for value in lst:
        if value in hits:
            hits[value] += 1
        else:
            hits[value] = 1

        highest = [value, hits[value]] if  highest == [] or hits[value] >
highest[1] else highest
    return highest
```

Standaard deviatie

De formule van de standaarddeviatie is als volgt:

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}}$$

Formule standaarddeviatie, van: https://en.wikipedia.org/wiki/Standard_deviation

De standaarddeviatie wordt gebruikt om het onderlinge verschil van waarden in een dataset aan te geven. De bovenstaande formule wordt vertaald naar Pythoncode in onderstaande snippet.

```
def standardDeviation(lst):
    avg = average(lst)
    quadrants = []
    for element in lst:
        quadrants.append((element - avg)**2)
    return average(quadrants)**.5
```

Lineaire regressie

Lineaire regressie wordt gebruikt om een lineaire lijn te kunnen trekken door verschillende punten, het doel is hierbij om de 'error' of het verschil tussen de punten en de lijn te minimaliseren. Hierbij wordt gebruikgemaakt van het gemiddelde van de x en de y- as van de lijn. Het product van de 2 lijsten van waardes van de x en de y-as worden hierna berekend. Hiervan wordt het product van de lengte van de x-as, het gemiddelde van de y-as, en het gemiddelde va de x-as afgetrokken. Vervolgens wordt hetzelfde gedaan maar dan met alleen 2 keer de x-as. Deze twee getallen worden door elkaar gedeeld. Het product van dit getal en het gemiddelde van de x-as wordt afgetrokken van de y-as, hiermee heb je de minimale error berekend.

```
def linearRegression(lst):
    x = list(range(1, length(lst) + 1))
    y = lst

    n = length(x)
    avg_x, avg_y = average(x), average(y)

    SS_xy = total([a*b for a,b in zip(y,x)]) - n*avg_y*avg_x
    SS_xx = total([a*b for a,b in zip(x,x)]) - n*avg_x*avg_x

    b_1 = SS_xy / SS_xx
    b_0 = avg_y - b_1*avg_x

    return x, y, (b_0, b_1)

y_pred = [b[0] + (e * b[1]) for e in x]
```

Betrouwbaarheidsinterval

Het betrouwbaarheidsinterval geeft aan in hoeverre je steekproef of dataset representatief kan zijn voor de hele populatie. Hiervoor geldt de volgende formule:

$$\mu_x = \bar{x} \pm t_{\alpha/2} * \frac{s_x}{\sqrt{n}}$$

Waarin:
- μ_x = verwachte gemiddelde in de populatie (de boven- en ondergrens worden berekend)
- $\bar{x}$ = het gevonden gemiddelde in de steekproef
- $t_{\alpha/2}$ = het betrouwbaarheidsinterval
- s_x = de standaarddeviatie van de steekproef
- n = het aantal respondenten

Formule van betrouwbaarheidsinterval, van: https://hulpbijonderzoek.nl/online-woordenboek/betrouwbaarheidsinterval/

In de code hieronder wordt uitgegaan van het 95% betrouwbaarheidsinterval. Volgens de centrale limietstelling wordt bij 95% het getal 1.96 gebruikt om in de formule te gebruiken.

```
def confidenceInterval(lst):
    std = standardDeviation(lst)
    l = length(lst)
    avg = average(lst)
    return [avg + 1.96 * (std / (l**.5)), avg - 1.96 * (std / (l**.5))]
```

GitHub

https://github.com/berryhijwegen/AC_HU/tree/master/statistiek/StatischeAnalyse

Bronnen

https://towardsdatascience.com/linear-regression-part-1-types-examples-gradient-descent-example-2e8c22b05f61

https://nl.wikipedia.org/wiki/Centrale_limietstelling

https://en.wikipedia.org/wiki/1.96

https://hulpbijonderzoek.nl/online-woordenboek/betrouwbaarheidsinterval/

https://en.wikipedia.org/wiki/Standard_deviation