Desktriptive Statistik

> For more help with python visit:
> http://www.scipy-lectures.org or take a course on https://www.datacamp.com/home.
>
> This summary was written with typora.

Ziele der deskriptiven Statistik

- Daten zusammenfassen durch **nummerische Kennwerte**
- **Graphische Darstellung** der Daten

Beispiel aus der deskriptiven Statistik

- **Bekannt**: n beobachtete Datenpunkte (Messungen) $x_1, x_2, \ldots, x_n$
- Wir berechnen die Lage- und Streuungsparameter und stellen diese graphisch dar (z.B. mit einem Boxplot)
- **Lageparameter**:
  - Arithmetisches Mittel (Durchschnitt / Schwerpunkt der Daten $\bar{x}_n$) $\rightarrow$ `seriesDataSet.mean()`
  - Median
  - Quantile
- **Streuungsparameter**:
  - Empirische Varianz / Standardabweichung
  - Quartilsdifferenz

Streuung

- Streuung nimmt Verteilung der Daten um den Mittelwert in Betracht
- Arithmetisches Mittel vernachlässigt diese Verteilung
- **Beispiel** Schulnoten einer Klasse (Arithmetisches Mittel)
  - Fall 1: Noten $\rightarrow$ 2, 6, 3, 5 ; Mittelwert $\rightarrow$ 4
  - Fall 2: Noten $\rightarrow$ 4, 4, 4, 4 ; Mittelwert $\rightarrow$ 4

Ansätze um die Streuung zu berechnen

- Es gibt drei verschiedene Ansätze um die Streuung zu berechnen. Wir verwenden den 3. Ansatz.

- **Ansatz 1**: Durchschnitt der Unterschiede zum Mittelwert
  - Fall 1: $\frac{(2-4)+(6-4)+(3-4)+(5-4)}{4} = 0$
  - Fall 2: $\frac{(4-4)+(4-4)+(4-4)+(4-4)}{4} = 0$
  - **Problem**: Unterschiede können negativ sein und sich gegenseitig auflösen

- **Ansatz 2**: Unterschiede durch Absolutwerte ersetzen (mittlere absolute Abweichung)
  - Fall 1: $\frac{|(2-4)|+|(6-4)|+|(3-4)|+|(5-4)|}{4} = 1.5 \rightarrow$ Noten weichen 1.5 vom Mittelwert
  - Fall 2: $\frac{|(4-4)|+|(4-4)|+|(4-4)|+|(4-4)|}{4} = 0$
  - **Problem**: Theoretische Nachteile

- **Ansatz 3**: Empirische Varianz $\rightarrow Var(x)$ und empirische Standardabweichung $\rightarrow s_x$
  - *"Für das Mass der Variabilität oder Streuung der Messwerte verwendet"*
  - Fall 1:

    $Var(x) =$ `seriesDataSetA.var()` $= 3.3$ $s_x =$ `seriesDataSetA.std()` $= 1.8257$
  - Fall 2: $Var(x) =$ `seriesDataSetB.var()` $= 0$ $s_x =$ `seriesDataSetB.std()` $= 0$

Empirische Varianz

- Kennzahl, um die Streuung eines Datensatzes zu beschreiben $\rightarrow$ `seriesDataSet.var()`
- Wenn empirische Varianz gross $\rightarrow$ Streuung um das arithmetische Mittel gross
- Hat keine physikalische Bedeutung

> Abweichungen $x_i - \bar{x}$ wird quadriert damit sich Abweichungen nicht gegenseitig aufheben können. Nenner $n - 1$ anstelle von n

Empirische Standardabweichung

- Kennzahl, um die Streuung eines Datensatzes **in derselben Einheit** zu beschreiben $\rightarrow$ `seriesDataSet.std()`
- Beispiel:
  - Anzahl Messungen $n = 13$
  - Arithmetisches Mittel $\bar{x_n} = 80.02 cal/g$
  - Empirische Varianz $Var(x) = 0.000574$
  - Standardabweichung $s_n = \sqrt{Var(x)} = 0.024 cal/g$
  - *„mittlere" Abweichung vom Mittelwert 80.02 cal/g ist 0.024 cal/g*

Median

- Lagemass für die "Mitte" → `seriesDataSet.median()`
- *"Wert, bei dem die Hälfte der Messwerte unter diesem Wert liegen"*
- **Berechnung**:
  1. Datensatz der Grösse nach sortieren
  2. Der **Median** ist nun der Wert mittleren Beobachtung (Messung) → aus 5 Beobachtungen ist der Median also die 3. Beobachtung
  3. Bei ungerader Anzahl Beobachtungen die mittlere Beobachtung nehmen
  4. Bei gerader Anzahl Beobachtungen den Durchschnitt der mittleren beiden Beobachtungen nehmen

Median vs. Arithmetisches Mittel

- Kommt auf die Problemstellung darauf an welches besser ist
- Am besten: beide Masse gleichzeitig verwenden
- Eigenschaften des Medians:
  - **robuster**, also
  - lässt sich weniger stark durch extreme Beobachtungen beeinflussen
  - (noch robuster wäre die Quartilsdifferenz (weiter unten))

Quartile

- Wert, wo `[Prozentsatz]` aller Beobachtungen `[kleiner oder gleich]` und `[1 - Prozentsatz]` `[grösser oder gleich]` sind wie dieser Wert
- Meistens existiert die `[Prozentsatz]`-igste Beobachtung nicht, dann müssen wir:
  - `[Prozentsatz]` der Anzahl Beobachtungen berechnen
  - Die erhaltene Zahl aufrunden und diese Beobachtung wählen (Zahl = 3.25, dann 4. Beobachtung wählen)
  - Falls die erhaltene Zahl gerade ist (z.B. 2), dann Durchschnitt von dieser Beobachtung und der nächsten Beobachtung als Quartil nehmen (2. und 3. Beobachtung)
- Python kennt nur Befehle für **Quantile**, aber nicht für **Quartile**
- Um **Quartile** zu berechnen geben wir die folgende Option in die `seriesDataSet.quantile()` - Funktion ein:
  - Unteres Quartil: `seriesDataSet.quantile(q=.25, interpolation="midpoint")`
  - Oberes Quartil: `seriesDataSet.quantile(q=.75, interpolation="midpoint")`

Unteres Quartil

- Wert, wo `25 %` aller Beobachtungen `kleiner oder gleich` und `75 %` `grösser oder gleich` sind wie dieser Wert

Oberes Quartil

- Wert, wo `75 %` aller Beobachtungen `kleiner oder gleich` und `25 %` `grösser oder gleich` sind wie dieser Wert

Quartilsdifferenz

- Kennzahl für die Streuung (Streuungsmass) der Daten
- $\text{oberes Quartil} - \text{unteres Quartil}$
- misst die **Länge des Intervalls**, das ca. die Hälfte der "mittleren" Beobachtungen enthält
- Je kleiner die Quartilsdifferenz, **umso näher liegt die Hälfte aller Werte um den Median**, also
- Kleinere Differenz, kleinere Streuung
- Dieses Streuungsmass ist **robust**