Statistische Datenanalyse

Joern Ploennigs · AI4SC

> Wir haben nur eine Welt, aber wir brauchen mehr als einen Weg, um sie zu verstehen.
>
> — Robert Merton

Nachdem man sich mit einem neuen Datensatz vertraut gemacht hat, werden grundlegende Statistiken berechnet, um die Daten besser zu verstehen, auf Plausibilität zu prüfen und wichtige Informationen für spätere Modellierungsschritte zu gewinnen, wie Verteilungsformen.

Einfache beschreibende Statistiken

Minimum, Maximum und der beobachtete Wertebereich

Wir explorieren in diesem Notebook den Wetterdatensatz aus Warnemünde und laden ihn mit:

```
import numpy as np # Import von NumPy
import pandas as pd # Import von Pandas

wetter = pd.read_csv("../data/Wetter/warnemuende_1960.csv", sep=';')
wetter.shape
```

```
(7209, 19)
```

Wetterdaten sind z.B. wichtig für die Planung von Betonierarbeiten, Trocknungszeiten und Bauzeitprognosen. Solche Informationen können auch helfen, Risiken und Verzögerungen besser zu kalkulieren.

Statistische Kennwerte wie Minimum, Maximum, Mittelwert und Standardabweichung helfen dabei, Daten zu normalisieren, Ausreißer zu erkennen oder Features richtig zu interpretieren. Das sind alles wichtige Schritte vor dem Trainieren von ML-Modellen.

Das Minimum und Maximum sind eine intuitive Statistik, die den kleinsten und größten Wert in einem Datensatz angeben. Wir können beide direkt bestimmen mit:

```
wetter.TMK.min(), wetter.TMK.max()
```

```
(np.float64(-14.8), np.float64(27.8))
```

Das Minimum und Maximum sind besonders sinnvoll zur Plausibilitätsprüfung des Datensatzes. Diese Werte zeigen, in welchem Bereich sich die Daten bewegen. Das ist z. B. wichtig, um unplausible Messwerte zu erkennen: etwa eine Temperatur von 120 °C im Winter. Hieraus lässt sich der beobachtete Wertebereich berechnen durch:

```
wetter.TMK.max() - wetter.TMK.min()
```

```
np.float64(42.6)
```

Der beobachtete Wertebereich ist ein sehr einfaches Maß für die Streuung in einem Datensatz, allerdings nicht sehr robust.

Achtung: Robustheit von Minimum und Maximum

Es ist wichtig zu beachten, dass das beobachtete Minimum, Maximum und der daraus folgende Wertebereich nicht den realen Grenzen entsprechen muss. Unter der Annahme das die Daten statistisch verteilt sind, liegen Minimum und Maximum häufig in Bereichen der Verteilung, die nur selten auftreten und folglich beobachtet werden. Es ist immer davon auszugehen, dass die realen Grenzwerte davon abweichen und einfach nur nicht beobachtet worden sind.

Nehmen wir als Beispiel die Temperatur. Aus der Physik ist bekannt, dass diese zwischen dem absoluten Nullpunkt bei −273,15 °C und der vermuteten maximalen Planck-Temperatur von 1,42x1032 °C liegen kann. Das ist der reale Wertebereich. Beide Temperaturen konnten allerdings noch nie gemessen werden und somit sollten sie in einem Messdatensatz nicht auftreten.

Mittelwert

Der arithmetische Mittelwert, oder Durchschnitt, ist ein beliebtes Maß für das Lagemaß. Mit dem Lagemaß beschreibt man die „typischen“ Werte in einem Datensatz. Sie folgen aus der_zentralen Tendenz_ vieler Wahrscheinlichkeitsverteilungen, dass die Werte um ein zentralen Wert liegen.

Der arithmetische Mittelwert entspricht der Summe aller Werte im Datensatz geteilt durch die Anzahl der Werte im Datensatz. Also, wenn wir Werte in einem Datensatz haben und sie Werte haben …, ist der Stichprobenmittelwert in der Regel durch den Buchstaben dargestellt. Die Formel ist:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Der Mittelwert ist ein zentrales Lagemaß des Datensatzes. Er beschreibt den durchschnittlichen Wert, muss aber nicht selbst als beobachteter Wert im Datensatz vorkommen. Eine wichtige Eigenschaft des Mittelwerts ist, dass er bei quadratischer Fehlerbewertung den Fehler bei der Vorhersage eines beliebigen Werts im Datensatz minimiert. Darüber hinaus ist der Mittelwert das einzige Lagemaß, bei dem die Summe der Abweichungen aller Werte vom Mittelwert immer 0 ist.

Der Mittelwert lässt sich in Pandas für jede Spalte mit der Methode `mean` direkt berechnen. Bestimmen wir für die Lufttemperatur in unserem Wetterdatensatz ergibt sich

```
wetter.TMK.mean()
```

```
np.float64(8.47518379803024)
```

Es war also im Schnitt 8.47°C.

Median

Der Mittelwert hat einen Hauptnachteil: Er ist besonders anfällig für den Einfluss von Ausreißern. Dies sind Werte, die im Vergleich zum Rest des Datensatzes ungewöhnlich sind, indem sie besonders klein oder groß im numerischen Wert sind. Betrachten wir zum Beispiel die Löhne der Mitarbeiter in einer Fabrik unten:

```
gehalt = [15, 18, 16, 14, 15, 15, 12, 17, 90, 95] # in Tausend
np.mean(gehalt)
```

```
np.float64(30.7)
```

Der durchschnittliche Gehalt für diese zehn Mitarbeiter beträgt `30,7` Tausend. Allerdings die Daten, dass dieser Durchschnittswert möglicherweise nicht der beste Weg ist, um das typische Gehalt eines Arbeitnehmers widerzuspiegeln, da die meisten Gehälter im Bereich von 12 bis 18 Tausend liegen. Der Mittelwert wird durch die beiden hohen Gehälter verzerrt. Daher braucht man einen *robusteres* Lagemaß für die zentrale Tendenz. Deshalb verwendet man in solchen Fällen meist den *Median*.

Ein weiterer Zeitpunkt, an dem wir in der Regel den Median gegenüber dem Mittelwert bevorzugen, ist, wenn unsere Daten schief sind (d. h. die Häufigkeitsverteilung unserer Daten ist schief). Wenn wir die Normalverteilung betrachten - da dies in der Statistik am häufigsten bewertet wird - sind der Mittelwert, der Median und der Modus identisch, wenn die Daten perfekt normal sind. Darüber hinaus stellen sie alle den typischsten Wert im Datensatz dar. Wenn die Daten jedoch schief werden, verliert der Mittelwert seine Fähigkeit, den besten zentralen Ort für die Daten zu liefern, da die schiefen Daten ihn vom typischen Wert wegziehen. Der Median behält jedoch am besten diese Position bei und wird nicht so stark von den schiefen Werten beeinflusst. Dies wird später im Abschnitt zur Schiefe genauer erläutert. In beiden Fällen nutzt man deshalb den Median.

Der Median ist der mittlere Wert für eine Reihe von Daten, die nach ihrer Größe geordnet wurden. Der Median wird weniger von Ausreißern und schiefen Daten beeinflusst.

Prinzipiell können wir ihn berechnen, indem wir die Werte zuerst der Größe nach sortieren:

```
gehalt.sort()
gehalt
```

```
[12, 14, 15, 15, 15, 16, 17, 18, 90, 95]
```

und dann den mittleren Wert auswählen, der sich aus der Länge berechnet

```
gehalt[len(gehalt) // 2]
```

```
16
```

Ist die Anzahl der Werte gerade, so bildet man den Durchschnitt aus den beiden in der Mitte liegenden Werten.

Der Einfachheit halber können wir den Median für eine Spalte direkt mit Pandas berechnen. Bei der Temperatur in dem Wetter Datensatz ergibt sich auch ein leicht unterschiedlicher Wert im Vergleich zum Mittelwert, nämlich

```
wetter.TMK.median()
```

```
np.float64(8.6)
```

Der mittelste Wert in unserem Datensatz war also 8.6°C und damit etwas wärmer als der Mittelwert von 8.47°C.

Standardabweichung und Varianz

Ein wichtiges Kriterium zur Beschreibung von Daten ist neben der zentralen Tendenz die Streuung der Daten. Der Grund ist, dass der Mittelwert oder der Median uns zwar eine Vorstellung des „typischen“ Wertes geben aber keine Information, wie stark die realen Werte davon abweichen.

In einem Datensatz mit einer großen Streuung ist der Mittelwert oder der Median nicht mehr sehr aussagekräftig. Wir haben das im Beispiel der Gehälter gesehen. Wo wir mit dem Median zwar wissen, dass diese um 16 Tausend liegen, aber dadurch auch ausblenden, dass die Gehälter bis zu 96 Tausend reichen.

Deshalb gibt man bei Daten meist die Varianz s^2 oder die Standartabweichung s die sich wie folgt berechnen

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

$$s = \sqrt{s^2}.$$

Die Formel zeigt, dass die Varianz s^2 das quadratische Abweichungsmaß der einzelnen Datenpunkte vom arithmetischen Mittel ist. Die Varianz ist ein Fehlermaß, wie weit die einzelnen Datenpunkte von diesem Mittelwert entfernt sind.

Die Standardabweichung s ist die Quadratwurzel der Varianz und gibt somit die durchschnittliche Abweichung der Datenpunkte vom Mittelwert in den ursprünglichen Einheiten der Daten an. Sie ist ein Maß für die Streuung der Datenpunkte um den Mittelwert. Die Standardabweichung ist oft intuitiver zu interpretieren, da sie direkt in den ursprünglichen Einheiten der Daten gemessen wird, während die Varianz in Quadraten dieser Einheiten gemessen wird.

In Pandas können wir beide Werte berechnen mit

```
wetter.TMK.var(), wetter.TMK.std()
```

```
(np.float64(48.71576448133568), np.float64(6.97966793489029))
```

Wir sehen an der Standardabweichung, dass die Temperatur im Durchschnitt um 6.9°C vom Mittelwert abweicht. Da die Varianz quadratisch ist, ist sie für Menschen schlechter zu interpretieren. Man benutzt sie dennoch zum Vergleichen, da durch die Quadratur, stärkere Abweichungen betont werden und sie so im Vergleich besser herausstechen. Eine Eigenschaft die später bei der Berechnung und der Bewertung von ML-Modellen relevant sein wird.

Schiefe

Die Schiefe ist ein Maß für die Asymmetrie der Datenverteilung. Ist die Schiefe 0 so sind die Werte symmetrisch um den Median verteilt. Ist die Schiefe negativ so streuen die Werte die kleiner als der Median sind mehr, was den arithmetischen Mittelwert auf diese Seite verschiebt. Bei einer positiven Schiefe streuen sie mehr auf der rechten Seite. Die Schiefe ist ein gutes Indiz für bestimmte Wahrscheinlichkeitsverteilungen.

Die Schiefe wird berechnet durch

$$g = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3.$$

Die Schiefe lässt sich auch in Pandas direkt berechnen. Für die Temperatur bestätigt sich jetzt unsere Beobachtung, die wir beim Vergleich von Median und Mittelwert gemacht haben, nämlich, dass die Verteilung etwas linksschief ist.

```
wetter.TMK.skew()
```

```
np.float64(-0.21288392100379444)
```

Wölbung

Der Wölbung oder Kurtosis ist ein Maß für die Spitzigkeit oder Flachheit einer Verteilung und ein Maß für die Abweichung von der Normalverteilung. Da viele ML-Modelle von normalverteilten Werten ausgehen, können Daten mit einer großen Kurtosis problematisch sein. Hier spricht man auch von Long-Tail-Verteilungen, also Verteilungen, die nicht schnell abflachen. Deshalb kann man sie im Data-Cleaning nicht gut abschneiden, um Außreißer zu entfernen und den Datensatz zu normalisieren, was wiederum numerische Problem mit sich führen kann. Spätestens wenn solche Probleme auftreten, sollte man sich die Kurtosis genauer ansehen.

Die Kurtosis wird berechnet durch

$$w = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4.$$

Vergleicht man die Formeln für Mittelwert, Varianz, Schiefe und Wölbung so zeigt sich ein Muster, das sich immer nur der Exponent ändert. Deshalb spricht man auch von dem ersten, zweiten, dritten und viertem statistischen Moment. Sie lassen sich auch durch eine gemeinsame Formel auf Basis des Erwartungswertes E beschreiben.

$$\mu_k = \mathrm{E}[(X - \mu)^k].$$

Für die Temperatur haben wir eine Kurtosis von `-0.7`. Das ist nicht problematisch. Solche Faustregeln für Schiefe und Kurtosis können eine erste Orientierung geben, ersetzen aber keine formale Prüfung auf Normalität. Sie zeigen hier lediglich, dass keine starke Abweichung von einer glockenförmigen Verteilung erkennbar ist [1], [2], [3].

```
wetter.TMK.kurt()
```

```
np.float64(-0.7045900367551634)
```

Quartile, IQR und Quartile & Perzentile

Die Diskussion der Robustheit des beobachteten Wertebereichs und des arithmetischen Mittelwertes, dass diese statistischen Größen anfällig sind gegenüber Ausreißern. Das trift auch auf die Varianz und die Standardabweichung zu. Gerade die quadratische Berechnung der Varianz überbetont solche Ausreißer. Deshalb braucht man Größen, die robuster reagieren, so genannte Quantile und Perzentile.

Die *Quartile* berechnen sich ähnlich zum der Median, bloß dass wir nicht den mittelsten Wert aus dem sortierten Datensatz wählen, sondern den Bereich in vier Teile teilen, wodurch das Q1-Quartile 25% der Daten kleiner sind, bei Q2-Quartile 50% der Daten kleiner sind (Median), und beim Q3-Quartile 75% der Daten drunter liegen. Wenn wir die Berechnungsvorschrift auf den Gehaltsdatensatz anwenden, so folgt:

```
gehalt = np.sort(gehalt)
Q1 = gehalt[len(gehalt) // 4 * 1]
Q2 = gehalt[len(gehalt) // 4 * 2]
Q3 = gehalt[len(gehalt) // 4 * 3]
Q1, Q2, Q3
```

```
(np.int64(15), np.int64(15), np.int64(17))
```

Dadurch das die 15 beim Q1 und Q2 vorkommt sehen wir schon, dass das ein häufiger Wert ist und aus dem 17 des Q3 folgt, dass der Datensatz rechtsschief ist.

Wenn wir den Datensatz nicht nur in vier Teile teilen sondern in 100 Teile, so, haben wir *Quantile*. Ein *p-Quantil* beschreibt, dass der Anteil p von 100 Teilen unter dem bestimmten Wert liegt. Es liegt nahe p in Prozent anzugeben und dann spricht man von *Perzentilen*. Nehmen wir zum Beispiel das 95%-Perzentil, so bedeutet das, dass 95 Prozent der Werte kleiner und 5 Prozent der Werte größer als das Perzentil sind. Damit sind Quantile auch eine generische Darstellung vom Quartilen (25%, 50%, 75%), Median (50%) und Minimum (0%) und Maximum (100%).

Perzentile lassen sich in Pandas wie gewohnt einfach berechnen, indem wir den gewünschten Prozentwert angeben.

```
wetter.TMK.quantile(0.25)
```

```
np.float64(3.0)
```

Wir können auch gleich mehrere Werte angeben.

```
t_min, t_01, t_Q1, t_median, t_Q3, t_99, t_max = wetter.TMK.quantile([0, 0.01,
0.25, 0.5, 0.75, 0.99, 1.0])
```

Hier bestimmen wir auch das 1%-Perzentil und das 99%-Perzentil. Solche Werte werden alternativ zum Minimum und Maximum verwendet, weil sie insbesondere bei größeren Datensätzen weniger betroffen von Ausreißern sind.

Als letztes statistische Maß betrachten wir den *Inter-Quartil-Abstand* (*IQR*). Er ist weniger empfindlich gegenüber Ausreißern als die Standardabweichung und wird alternativ als robustes Streuungsmaß verwendet. Der IQR misst die Spannweite der mittleren 50% der Daten, oder anders gesagt die Differenz des Q3-Quartils bei 75% und des Q1-Quartils bei 25%.

```
t_IQR = t_Q3 - t_Q1
t_IQR
```

```
11.5
```

Aufgrund seiner robusten Eigenschaften wird der IQR gerne zur Identifikation von Ausreißern benutzt. Dabei nimmt man an das Werte die mehr als 1,5 IQR (oder 3 IQR wenn man vorsichtig ist) vom 1. oder 3. Quantil entfernt sind, als Ausreißer angesehen werden können. Wenn wir das für die Temperatur bestimmen, so ergibt sich ein Ausreißer von `-14.8` °C der 1956 gemessen worden ist.

```
wetter[(wetter.TMK < t_Q1 - 1.5 * t_IQR) | (wetter.TMK > t_Q3 + 1.5 * t_IQR)]
```

	STATIONS_ID	MESS_DATUM	QN_3	FX	FM	QN_4	RSK	RSKF	SDK	SHK_TAG	NM	VPM	PM	TMK	UPM	TXK	TNK	TGK	eor
3328	4271	19560215	5	−999	1.9	5	0.0	8	4.3	20	2.4	1.9	1010.0	−14.8	91.0	−10.1	−16.6	−24.0	eor

Viele von den bisher besprochenen Kenngrößen lassen sich in Pandas auch mit einer einzigen Funktion für alle numerischen Spalten bestimmen, so dass man es sich sparen kann die ganzen Funktionen auf jede Spalte anzuwenden:

```
wetter.describe()
```

	STATIONS_ID	MESS_DATUM	QN_3	FX	FM	QN_4	RSK	RSKF	SDK	SHK_TAG	NM	VPM	PM	TMK	UPM	TXK	TNK	TGK
count	7209.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
mean	4271.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
std	0.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
min	4271.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
25%	4271.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
50%	4271.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
75%	4271.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
max	4271.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Verteilungsformen

Die Kenntnis verschiedener statistischer Verteilungen ist für die Datenanalyse von Bedeutung. Wenn man die Verteilungsform und ihre Parameter von Variablen kennt, so kann man auf die Entstehungsprozesse schließen und die Daten auch synthetisch erzeugen, was z.B. wichtig für Simulationen ist.

Kontinuierliche Verteilungen

Normalverteilung

Die Normalverteilung, auch Gaußsche Glockenkurve genannt, ist eine symmetrische Verteilung mit einem Maximum in der Mitte und abnehmenden Wahrscheinlichkeiten in den Flanken. Sie ist

die am häufigsten vorkommende Verteilung in der Natur. Das ist unter anderem auf den zentralen Grenzwertsatz zurückzuführen:

📘 Definition: Zentrale-Grenzwertsatz

Der zentrale Grenzwertsatz von Lindeberg-Lévy besagt, dass wenn wir viele unabhängige Zufallsvariablen addieren, unabhängig von ihrer ursprünglichen Verteilung, die Summe dieser Zufallsvariablen tendenziell einer Normalverteilung folgt, wenn die Anzahl der Zufallsvariablen groß genug ist.

Infolgedessen findet sich die Normalverteilung bei vielen Größen, die von vielen Einflussgrößen beeinflusst werden, unteranderem die Körpergröße von Menschen oder der Messfehler bei Messungen, da die zufälligen Fehler von Messgeräten viele Ursachen haben und sich meist zu einer Normalverteilung überlagern.

Die Normalverteilung hat als charakteristische Parameter den Mittelwert μ und die Standardabweichung σ. Sie ist definiert durch die folgende Funktion

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Zu erkennen ist die Normalverteilung durch die charakteristische Glockenform.

```
Unable to display output for mime type(s): text/html
```

```
Unable to display output for mime type(s): text/html
```

Gleichverteilung
Bei der Gleichverteilung ist die Wahrscheinlichkeit für alle Ereignisse gleich über einen festen Bereich, außerhalb dieses Bereiches ist sie null. Die Gleichverteilung kann sowohl kontinuierlich als auch diskret sein. In der Natur ist sie zum Beispiel beim Würfeln zu beobachten. Auch Computer erzeugte Zahlen sind meist Gleichverteilt.

Die Bereichsgrenzen a und b sind die Parameter der Gleichverteilung die definiert ist als:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for} \quad a \leq x \leq b, \\ 0 & \text{for} \quad x < a \quad \text{or} \quad x > b. \end{cases}$$

Zu erkennen ist die Geleichverteilung an der kastenähnlichen Verteilung.

```
Unable to display output for mime type(s): text/html
```

Exponentialverteilung
Bei der Exponentialverteilung sind die Auftrittswahrscheinlichkeiten durch die Exponentialfunktion beschreibbar. Hierbei ist die Wahrscheinlichkeit, dass ein Wert 0 oder klein ist, deutlich größer als das er groß ist. Die Exponentialverteilung tritt oft bei der Modellierung von Zeiten

bis zum Eintritt von Ereignissen auf, die im Durchschnitt mit konstanter Rate eintreten, z.B. die Lebensdauer von Atomen bei radioaktiven Zerfällen oder die Wartezeiten zwischen Anrufen in einem Call-Center.

Die Exponentialverteilung ist durch die negative Exponentialfunktion definiert und kann damit nur positive Werte annehmen. Der charakteristische Parameter λ wird oft auch als Auftrittsrate oder Bearbeitungsrate bezeichnet. Formal ist die Verteilung definiert durch

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Zu erkennen ist die Verteilung an der Exponentialfunktion und der hohen Wahrscheinlichkeit von Werten die 0 oder fast 0 sind. Ferner ist sie typisch für Variablen die Zeiten zwischen Ereignissen modellieren, die mit konstanter Rate auftreten.

```
Unable to display output for mime type(s): text/html
```

Pareto-Verteilung

Die Pareto-Verteilung, benannt nach Vilfredo Pareto, gleicht im ersten Eindruck der Exponentialverteilung. Sie wird oft verwendet, um Phänomene zu modellieren, bei denen eine kleine Anzahl von Elementen einen Großteil der Gesamtwerte ausmacht. Sie wird häufig in wirtschaftlichen und sozialen Studien verwendet, um die Verteilung von Einkommen, Vermögenswerten oder anderen Ressourcen zu beschreiben. (z.B. das 20/80-Prinzip: 20% der Arbeit führen zu 80% der Ergebnisse).

Die Wahrscheinlichkeitsdichte der Pareto-Verteilung basiert nicht auf der Exponentialverteilung sondern ist proportional zu $1/x^k$. Sie wird bestimmt durch den Offset $x_{\min} > 0$ und der Steigung $k > 0$.

$$f(x) = \begin{cases} \frac{k x_{\min}^k}{x^{k+1}} & x \geq x_{\min}, \\ 0 & x < x_{\min}. \end{cases}$$

Die Verteilung sieht erstmal der Exponentialverteilung ähnlich, hat aber eine andere, flexiblere Form, die durch die zwei Parameter bestimmt ist. Sie liegt meist vor bei Werten, bei denen einige wenige Elemente einen Großteil der Gesamtwerte ausmachen.

```
Unable to display output for mime type(s): text/html
```

Gammaverteilung

Die Gammaverteilung ist eine kontinuierliche Wahrscheinlichkeitsverteilung, die sowohl der Exponentialverteilung gleichen kann als auch der Normalverteilung. Die Gammaverteilung wird häufig verwendet, um die Summe von unabhängigen Exponentialverteilungen zu modellieren. Sie kann als eine verallgemeinerte Exponentialverteilung angesehen werden. Sie wird häufig verwendet, um die Wartezeit bis zum Auftreten eines Ereignisses zu modellieren, sowie in verschiedenen anderen Anwendungen.

Die Parameter der Gammaverteilung ist der Formparameter $\alpha > 0$ und der Skalen- oder Ratenparamter $\beta > 0$. Die Verteilung ist positiv definiert durch

$$f(x) = \{\frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)}, x \geq 0,$$
$$0, x < 0.$$

Die Gammaverteilung kann man von der Exponentialverteilung durch linksseitigen Anstieg unterscheiden. Die Gammaverteilung kann rechtschief und linksschief sein. Wenn die Summe von Exponentialverteilungen modelliert werden soll, ist die Gammaverteilung die richtige Wahl.

```
Unable to display output for mime type(s): text/html
```

Weibul-Verteilung
Die Weibull-Verteilung wird häufig verwendet, um die Lebensdauer von Materialien und Bauteilen zu modellieren. Sie findet auch Anwendung in der Meteorologie, der Hydrologie und der Zuverlässigkeitstechnik und Ausfallzeitanalyse.

Die Weibull-Verteilung ist eine Verteilung mit Formparameter $k > 0$ und Skalierungsparameter $\lambda > 0$. Ein wichtiger Spezialfall ist die Exponentialverteilung, die sich für $k = 1$ ergibt. Formal ist sie definiert durch

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Für viele Parameterwerte ist die Weibull-Verteilung rechtsschief. Wenn die Lebensdauer von Materialien oder Bauteilen modelliert werden soll, ist die Weibull-Verteilung oft die richtige Wahl.

```
Unable to display output for mime type(s): text/html
```

Betaverteilung
Die Betaverteilung ist eine Familie von kontinuierlichen Wahrscheinlichkeitsverteilungen auf dem Intervall $[0, 1]$, die durch zwei positive Parameter kontrolliert wird. Sie wird oft in der Bayes-Statistik verwendet.

```
Unable to display output for mime type(s): text/html
```

Diskrete Verteilungen

Binomialverteilung
Die Binomialverteilung beschreibt die Wahrscheinlichkeit einer bestimmten Anzahl von Erfolgen in einer Reihe von unabhängigen Versuchen mit fester Erfolgswahrscheinlichkeit. Sie wird z.B. verwendet, um die Anzahl der „Köpfe“ beim Münzwurf zu modellieren.

Sie ist definiert mit der Anzahl der Versuche $n > 0$ und der Erfolgs- oder Trefferwahrscheinlichkeit $p \in [0, 1]$ als Parameter zu

$$P(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{falls} \quad k \in \{0, 1, ..., n\}, \\ 0 & \text{sonst.} \end{cases}$$

```
Unable to display output for mime type(s): text/html
```

Poisson-Verteilung
Die Poisson-Verteilung beschreibt die Wahrscheinlichkeit einer bestimmten Anzahl von Ereignissen in einem bestimmten Zeitintervall oder einer bestimmten Fläche. Sie wird z.B. verwendet, um die Anzahl der Anrufe in einem Callcenter pro Stunde zu modellieren.

Die Poisson-Verteilung ist beschrieben durch die Ereignishäufigkeit λ zu

$$P(k) = \frac{\lambda^k}{k!} \, \mathrm{e}^{-\lambda}.$$

```
Unable to display output for mime type(s): text/html
```

Referenzen

Bibliographie

[1] B. M. Byrne, *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. routledge, 2013.

[2] R. B. Kline, *Principles and practice of structural equation modeling*. Guilford publications, 2023.

[3] J. F. Hair Jr, W. C. Black, B. J. Babin, und R. E. Anderson, „Multivariate data analysis“, *Multivariate data analysis*. S. 785, 2010.