
Datenanalyse in der Physik

Spring Semester 2026

Lecture Notes

Contents

1 Einführung in Python	1
1.1 Ein Erstes Python Programm	1
2 Die Messung als Zufallsprozess	2
2.1 Fehlerfortpflanzung	3
2.2 Korrelationen	4

References

- [1] Vorlesungsskript

Datenanalyse in der Physik

Fynn Krebser—fkrebser@student.ethz.ch

Spring Semester 2026

Lec 1

1 Einführung in Python

Wir verwenden Python als Programmiersprache, unter anderem, da Python sehr populär ist. Ein Grund dafür ist, dass Python einfach zu lernen und zu lesen ist. Weiter ist Python interaktiver als zum Beispiel C++. Als Programmierungsumgebung in diesem Kurs verwenden wir Jupyter Notebooks.

1.1 Ein Erstes Python Programm

Betrachten wir das folgende Python Programm:

```
def compute_mean(measurements):  
    """Compute mean of a list of measurements."""  
  
    if len(measurements) == 0:  
        return 0  
  
    sum = 0  
    for value in measurements:  
        sum += value  
    mean = sum / len(measurements)  
    return mean
```

```
measurements = [1, 0.1, 2.0]  
compute_mean(measurements)
```

Kommentare können auf Zwei Arten geschrieben werden. Einzeilige Kommentare beginnen mit einem # und mehrzeilige Kommentare werden von drei Anführungszeichen umschlossen.

In Python haben Variablen keinen Typ. Dafür müssen Variablen initialisiert werden. Der Typ wird dann automatisch bestimmt.

```
a = 3  
a/2
```

In diesem Beispiel erhalten wir 1.5.

Weiter wichtig in Python das Codeblöcke durch Einrückungen definiert werden.

Das if/else statement in Python sieht wie folgt aus:

```
if condition:  
    # do something  
elif condition2:  
    # do something else  
else:  
    # do something else
```

Ähnlich funktionieren while Schleifen:

```
x = 0  
while x < 5:  
    print(x)  
    x = x+1
```

Eine for Schleife sieht wie folgt aus:

```
for x in range(0,5):  
    print(x)
```

Hierbei erzeugt range(0,5) eine Liste mit den Zahlen 0,1,2,3,4.

In Python können Listen unterschiedliche Typen besitzen und sehen wie folgt aus:

```
l = [3, "Hund", "Katze"]
```

Die Indexierung ist analog zu C++ 0-basiert. Ausserdem hat das letzte Element Index -1.

Eine hilfreiche Funktion kann die help Funktion sein. Diese gibt die Dokumentation eines Objekts zurück. Zum Beispiel:

```
help(l)
```

gibt die Dokumentation der Liste zurück.

In Python ist ALLES eine Referenz. Wenn wir also b=a setzen und a bearbeiten, dann wird auch b bearbeitet.

Auch wichtig sind Tupel. Diese sind ähnlich wie Listen, aber unveränderlich. Sie werden mit runden Klammern definiert:

```
t = (1, 2, 3)
```

Die Klammern sind optional, aber es ist eine gute Praxis sie zu verwenden.

Tupel unpacking ist eine nützliche Funktion, die es ermöglicht, die Werte eines Tupels in einzelne Variablen zu entpacken. Zum Beispiel:

```
number, name = (1, "Hund")
```

Funktionen werden mit dem def Schlüsselwort definiert, wie im Einführungsbeispiel. Um mehrere Werte zurückzugeben, kann ein Tupel verwendet werden.

Sinnvolle Wissenschaftliche Bibliotheken beinhalten unter anderem: NumPy, SciPy, Pandas und Matplotlib.

Ein Array wird in NumPy wie folgt erstellt:

```
array_1d = np.array([1, 2, 3])  
array_2d = np.array([[1, 2, 3], [4, 5, 6]])  
array_range = np.arange(0,4, step=0.5)
```

Mathematische Operationen können direkt auf Arrays angewendet werden:

```
np.square(array_1d)  
np.sqrt(array_1d)
```

Um Matplotlib zu importieren, verwenden wir:

```
import matplotlib.pyplot as plt
```

2 Die Messung als Zufallsprozess

Ein Modell ist eine vereinfachte Vorstellung welches nie vollständig und nicht bewiesen werden kann. Mit einem Experiment können wir es lediglich falsifizieren.

Ein Experiment hingegen ist nie fehlerfrei und auch die Interpretation ist nicht eindeutig.

Betrachten wir als Modell periodische Funktionen. Hierbei ist $\phi(t)$ die Phase und es gilt

$$\tan(\phi) = \frac{\Im(z)}{\Re(z)}.$$

Mit einer linearen zeitabhängigen Phase erhalten wir eine Oszillation der Form

$$x(t) = x_0 \cdot \Re(e^{i\phi(t)}).$$

Der getriebene gedämpfte harmonische Oszillator erfüllt die Gleichung:

$$\ddot{x} + \Gamma \dot{x} + \omega_0^2 x = \frac{F_0 \cos(\omega t)}{m}.$$

Nach einer langen Zeit erreicht die Lösung die Form

$$x_0(\omega) = \left| \frac{F_0/m}{\omega_0^2 - \omega^2 + i\Gamma\omega} \right|.$$

Wenn wir einen Quarzkristall haben, dann sieht die Resonanzkurve wie folgt aus: INSERT FIGURE

Der unterschied zur erwarteten Kurve entsteht dadurch, dass die beiden Platten einen Kondensator bilden, welcher die Resonanzfrequenz verschiebt. somit ist

$$I_{\text{in}}(\omega) \propto |x(\omega) + K(\omega)| \neq x_0.$$

Somit haben wir einen systematischen Fehler begangen. Das Modell war nicht vollständig. Mit einem besseren Modell können wir diesen Fehler korrigieren.

Definition 2.1: Systematischer Fehler

Ein Fehler, welcher durch ein unvollständiges Modell entsteht, wird als systematischer Fehler bezeichnet. Er tritt bei jeder Messung auf und kann durch ein verbessertes Modell korrigiert werden.

Bei einer Messung können auch zufällige Fehler auftreten, was als **RAUSCHEN** bezeichnet wird. Wir simulieren uns einen Analog-Digital-Converter (ADC). Hierbei gibt es eine zeitliche Auflösung, eine maximal Messbare Spannung und eine Anzahl von Bits, welche die Auflösung der Messung bestimmt.

Wir definieren folgende Größen:

- Zeitlicher Abstand zwischen Messungen: Δ_t
- Sampling Rate / Abtastrate: $f_{\text{sample}} = 1/\Delta_t$
- Bandbreite der Messung: $f_{\text{bandwidth}} = \frac{1}{2\Delta_t}$
- Totale Messzeit: t_{tot}
- Spektrale Auflösung: $\Delta f = 1/t_{\text{tot}}$
- Signalauflösung: U_{min}

- Signalbereich: U_{max}
- Rauschen der Messapparatur: U_{noise}

Die Meisten dieser Dinge führen zu statistischen Fehlern, welche durch wiederholte Messungen korrigiert werden können.

Wenn wir eine Grösse messen mit erwartetem Wert $\tilde{x}$ und einen Wert x_1 messen, dann bezeichnen wir den **FEHLER** als $e_1 = \tilde{x} - x_1$.

Das Problem ist, dass wir a priori $\tilde{x}$ nicht kennen.

Die Möglichen Fehler können wie bereits erwähnt in systematische und statistische Fehler unterteilt werden. Systematische Fehler können durch eine Kontrollmessung korrigiert werden. Statistische Fehler können nicht verhindert werden. Wir können sie aber durch wiederholte Messungen quantifizieren.

Um einen gegebenen Datensatz zu analysieren, diskretisieren wir den Messbereich x in gleich grosse **BINS**.

Dies kann man dann in einem Histogramm darstellen.

Diese Daten sind zufällig verteilt um einen Mittelwert mit einer Streuung.

Definition 2.2: Mittelwert

Der Mittelwert eines Datensatzes $x_1, x_2, \dots, x_N$ ist definiert als

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

Um die Streuung um den Mittelwert zu quantifizieren, verwenden wir die Standardabweichung:

Definition 2.3: Varianz und Standardabweichung

Die Varianz eines Datensatzes $x_1, x_2, \dots, x_N$ ist definiert als

$$\text{var}(x) = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2.$$

Die Standardabweichung ist die Quadratwurzel der Varianz:

$$s = \sqrt{\text{var}(x)}.$$

Achtung: Die Standardfunktionen von Python verwenden die Population Varianz, welche durch N dividiert wird, anstatt durch $N-1$.

Wir definieren weiter die **GENAUIGKEIT** einer Messung als die Abweichung des Mittelwerts von dem wahren Wert $\tilde{x}$.

Ausserdem ist die **PRÄZISION** einer Messung definiert als das Doppelte der Standardabweichung. Sie ist ein Mass für zufällige Fehler.

Es existieren verschiedene Wahrscheinlichkeitsverteilungen, wobei die **NORMALVERTEILUNG** die wichtigste ist. Dies ist aufgrund des zentralen Grenzwertsatzes der Fall, welcher besagt, dass die Summe von vielen unabhängigen Zufallsvariablen, welche eine endliche Varianz haben, normalverteilt ist.

Die Normalverteilung ist eine kontinuierliche Wahrscheinlichkeitsdichtefunktion, welche durch die folgenden Parameter definiert ist:

$$PDF(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Bei einer Normalverteilung ist die Wahrscheinlichkeit, dass eine Messung innerhalb von einem Standardabweichung vom Mittelwert liegt, etwa 68%, innerhalb von zwei Standardabweichungen etwa 95% und innerhalb von drei Standardabweichungen etwa 99.7%.

Mit diesen Werten kann effizient geprüft werden, ob die Daten normalverteilt sind.

Unsere Konvention ist es für die empirischen Größen, $\bar{x}$ und Δx zu verwenden, während die Werte für die Normalverteilung mit $\mu = E(x) = \langle x \rangle$ und σ bezeichnet werden.

Zusammenfassend können wir einen Messwert angeben durch

$$x_m = \bar{x} \pm \Delta x.$$

Diese Aussage ist nur eine Angabe für die Messung eines Datenpunkts. Es ist keine Aussage über die Genauigkeit von $\bar{x}$.

Lec 4

Wir wollen nun den Mittelwert abschätzen. Wenn wir eine Messung wiederholt durchführen, so wird die Abschätzung des Mittelwerts immer genauer. Die Standardabweichung bleibt aber erhalten. Der Fehler in der Standardabweichung wird jedoch auch genauer.

Der Mittelwert ist definiert als $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$. Für $N \rightarrow \infty$ konvergiert $\bar{x}$ gegen den wahren Wert μ . Der Wert von $\bar{x}$ ändert sich nicht systematisch mit N , sondern schwankt um μ herum. Wenn wir die Varianz des Mittelwerts berechnen, so erhalten wir

$$\text{var}(\bar{x}) = \frac{1}{N^2} \sum_{i=1}^N \text{var}(x_i) = \frac{\sigma^2}{N}.$$

Wobei wir verwendet haben, dass $\text{var}(\sum x_i) = \sum \text{var}(x_i)$, da die x_i unkorreliert sind. Somit ist die Standardabweichung des Mittelwerts

$$\Delta \bar{x} = \frac{\sigma}{\sqrt{N}}.$$

Die Standardabweichung des Mittelwerts kann somit kleiner als die Standardabweichung der Daten sein.

Zusammenfassend, μ und σ ändern sich nicht systematisch mit N , aber deren Abschätzungen werden genauer, mit N .

2.1 Fehlerfortpflanzung

Häufig können wir nicht direkt die Grösse messen, welche wir interessiert, sondern müssen sie aus anderen Messgrößen berechnen.

Zum Beispiel hängt die Amplitude eines Pendels ab von

$$A = f(m, g, T, \dots).$$

Welchen Einfluss haben die Fehler in m , g und T auf den Fehler in A ?

Um das Problem zu vereinfachen, betrachten wir $f(x)$ mit einem Fehler. Wir können $f(x)$ um den Mittelwert $\bar{x}$ entwickeln:

$$f(x) = f(\bar{x} + \delta x) = f(\bar{x}) + \delta x f'(\bar{x}) = \bar{f} + \delta f(x).$$

Für den Fall der Standardabweichung von f erhalten wir

$$f(x) \approx f(\bar{x}) + \left. \frac{\partial f}{\partial x} \right|_{\bar{x}} (\Delta x) = \bar{f} + \Delta f(x).$$

Für zwei Variablen x und y erhalten wir

$$\begin{aligned} \Delta f^2(x, y) &= \frac{1}{N-1} \sum_{n=1}^N (f_n - \bar{f})^2 \\ &= \frac{1}{N-1} \sum_{n=1}^N \left(\frac{\partial f}{\partial x} (x_n - \bar{x}) + \frac{\partial f}{\partial y} (y_n - \bar{y}) \right)^2 \\ &= \left(\frac{\partial \bar{f}}{\partial x} \right)^2 + \left(\frac{\partial \bar{f}}{\partial y} \right)^2 \\ &\quad + 2 \frac{\partial \bar{f}}{\partial x} \frac{\partial \bar{f}}{\partial y} \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y}). \end{aligned}$$

Der Erste Term ist das Resultat für unkorrelierte Variablen. Der Zweite Term ist die Korrelation zwischen x und y . Wenn x und y unkorreliert sind, dann verschwindet der zweite Term.

Dies kann geschrieben werden als

$$\sigma_f^2 = (\partial_x, \partial_y) f \begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{pmatrix} \begin{pmatrix} \partial_x f \\ \partial_y f \end{pmatrix}.$$

Für unkorrelierte Variablen, also $\sigma_{xy}^2 = 0$, erhalten wir

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x} \right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y} \right)^2 \sigma_y^2.$$

Die Gauss-Methode setzt voraus, dass die Fehler klein sind, so dass die lineare Näherung gültig ist.

Nehmen wir als Beispiel ein Federpendel mit $T = 2.5s \pm 0.08s$ und $m = 0.3kg \pm 0.02kg$.

Die Federkonstante ist gegeben durch

$$k = m \left(\frac{2\pi}{T} \right)^2 \Rightarrow \bar{k} = 1.89 \frac{N}{m}.$$

Wir bilden die partiellen Ableitungen:

$$\frac{\partial k}{\partial m} = \left(\frac{2\pi}{T} \right)^2 = 6.32 \frac{N}{m \cdot kg}.$$

$$\frac{\partial k}{\partial T} = -4\pi m \left(\frac{2\pi}{T^3} \right) = -0.96 \frac{N \cdot s}{m}.$$

$$\Delta k = \sqrt{\left(\frac{\partial k}{\partial m} \right)^2 \Delta m^2 + \left(\frac{\partial k}{\partial T} \right)^2 \Delta T^2} = 0.175 \frac{N}{m}.$$

Es macht keinen Sinn mehr Stellen anzugeben, als die Fehler erlauben.

$$k = (1.9 \pm 0.2) \frac{N}{m}.$$

Tip 2.4:

Man sollte die Fehlerrechnung sowie korrektes Runden an der Prüfung können.

2.2 Korrelationen

Wir erinnern uns, dass die Fehlerfortpflanzung ein Korrelations-Term enthalten war. Damit wollen wir uns nun beschäftigen.

Stellen wir uns ein Pendel vor, mit einer Resonanzfrequenz f und der Temperatur T . Wir können nun die Resonanzfrequenz über einen Tag messen. Wenn wir die Temperatur gleichzeitig messen, so können wir erkennen, dass beide Kurven sehr einen ähnlichen Verlauf haben. Es gibt eine sogenannte **KORRELATION** zwischen f und T .

Wie kann nun Korrelation anschaulich erklärt, und mathematisch formuliert werden?

Eine Korrelation bedeutet, dass eine Änderung in T mit einer Änderung in f einhergeht. Mathematisch beschreiben wir die Variable T als $T = \bar{T} + \delta T$ und die Variable f als $f = \bar{f} + \delta f$. Der Ausdruck

$$(x_n - \bar{x})(y_n - \bar{y}),$$

sollte entweder immer positiv oder immer negativ sein.

Formell definieren wir

Definition 2.5: Kovarianz

Die Kovarianz zwischen zwei Variablen x und y ist definiert als

$$\sigma_{xy}^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y}).$$

Visuell möchten wir sagen, dass eine Korrelation stark ist, wenn sie grösser als die Standardabweichungen der Variablen ist. Daher definieren wir den

Definition 2.6: Korrelationskoeffizient

Der Korrelationskoeffizient zwischen zwei Variablen x und y ist definiert als

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \in [-1, 1].$$

Diese Grösse ist nun Normiert, wenn die Korrelation stark ist, dann ist $\rho_{xy} \approx 1$ oder $\rho_{xy} \approx -1$. Wenn die Korrelation schwach ist, dann ist $\rho_{xy} \approx 0$.

Im ersten Fall sprechen wir von maximaler (anti-) Korrelation, im zweiten Fall von keiner Korrelation.

Betrachten wir nun die nichtlineare Funktion $y = x^2$, wobei x eine Zufallsvariable mit symmetrischer Verteilung um 0 ist. In diesem Fall erhalten wir für die Kovarianz $\sigma_{xy}^2 = 0$, obwohl y von x abhängt.

Mit dem Korrelationskoeffizienten können wir somit nur auf lineare Korrelationen testen. Nichtlineare Fälle müssen generell anders behandelt werden.

Korrelation kann auch implizieren dass wir ein sehr grosses Rauschen haben, welches aber gleichzeitig beide Variablen beeinflusst. Wichtig, von Korrelation können wir nicht auf Kausalität schliessen. Zum Beispiel korrelieren die Anzahl der Piraten mit der globalen Durchschnittstemperatur.

Wir betrachten die Funktion $x = x^{(1)} + x^{(2)}$, wobei $x^{(1)}$ und $x^{(2)}$ jeweils N mal gemessen wurden.

Für die Varianz von x erhalten wir

$$\text{var}(x) = \text{var}(x^{(1)}) + \text{var}(x^{(2)}) + 2 \text{cov}(x^{(1)}, x^{(2)}).$$

Generell gilt

$$\begin{aligned} \text{var} \left(\sum_{i=1}^N a^{(k)} x^{(k)} \right) \\ = \sum_{i=1}^N a^{(k)2} \text{var}(x^{(k)}) + 2 \sum_{i < j} a^{(i)} a^{(j)} \text{cov}(x^{(i)}, x^{(j)}), \end{aligned}$$

wobei $a^{(k)}$ Konstanten sind.

Wenn wir in die Kovarianz für $y = x$ einsetzen, so erhalten wir

$$\text{cov}(x, x) = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x}) = \text{var}(x).$$

Wenn wir nun aber für $y = x + \Delta$ einsetzen, also zum Beispiel eine zeitliche Verschiebung, so erhalten wir

$$R_{xx}(\Delta) = \frac{1}{N-1-\Delta} \sum_{n=1}^{N-\Delta} (x_n - \bar{x})(x_{n+\Delta} - \bar{x}).$$

Dies wird die **DISKRETE AUTO-KOVARIANZ** genannt. Eine Andere Schreibweise wäre

$$R_{xx}(\tau) = \frac{1}{N-1-\Delta} \sum_{t=t_1}^{t_N-\tau} (x(t) - \bar{x})(x(t+\tau) - \bar{x}).$$

Wir können analog zur Kovarianz auch den Autokorrelationskoeffizienten definieren:

$$\rho_{xx}(\tau) = \frac{R_{xx}(\tau)}{R_{xx}(0)} = \frac{R_{xx}(\tau)}{\text{var}(x)}.$$

Weniger relevant für echte Daten, aber dennoch interessant, ist die kontinuierliche Autokovarianz:

$$R_{xx}(\tau) = \frac{1}{T} \int_0^T (x(t) - \bar{x})(x(t+\tau) - \bar{x}) dt.$$

Wenn ein Signal in der Autovarianz-basis Periodisch ist, so ist es auch in der Zeitbasis periodisch.

Die Auto-Kovarianz ist symmetrisch um $\tau = 0$. Die Samplingzeit setzt ein unteres Limit für die zeitliche Auflösung der Autokovarianz.

Index

Bins, 2
diskrete Auto-
Kovarianz, 4
Fehler, 2
Genauigkeit, 2
Korrelation, 4
Normalverteilung, 2
Präzision, 2
rauschen, 2