# Normalverteilung: und bestimmen

In diesem Blog wird es sehr mathematisch. Wir werden eine Menge ableiten und herleiten. Wenn Du also Freude am Formelnreiten hast, solltest Du Dich an die gute alte Zeit aus dem Mathe LK erinnert fühlen. Mir hat diese Fingerübung so viel Spaß bereitet, dass ich beschlossen habe, diesen kleinen Blog zu schreiben.

## Die Normalverteilung

Maschinelles Lernen braucht viele Daten, um den richtigen Algorithmus erlernen zu können. Wenn man zum Beispiel ein ML Verfahren entwickelt, dass das Wetter der nächsten Tage bestimmen soll, braucht man möglichst viele Messwerte (Datum, Temperatur, Feuchtigkeit, Luftdruck, …) aus der Vergangenheit. Natürlich darf man den Daten, mit denen man seinen Algorithmus füttert nicht blind vertrauen. Man muss sich auf jeden Fall seiner Qualität vergewissern, um nicht etwas Falsches zu lernen. Da man sich zumeist nicht jeden Datensatz einzeln anschauen kann, verwendet man statistische Verfahren, um etwaige Ausreißer und Fehler herausfiltern zu können.

Die Normalverteilung ist eins der wesentlichen Werkzeuge eben dafür. Dabei geht es darum, einen Ausdruck zu finden, wie wahrscheinlich ein durchschnittliches Ergebnis ist. Eine Frau in Deutschland ist zum Beispiel durchschnittlich 1,66m groß. Aber natürlich sind nicht alle Frauen in Deutschland 1,66m groß. Die Normalverteilung kann für diesen Fall eine quantitative Aussage treffen, wie groß die Wahrscheinlichkeit ist, dass die nächste Frau, die Du siehst 1,66m, 1,80m oder 1,25m groß ist. Die Normalverteilung gehorcht dabei folgender Gleichung:

Die Fläche unter dieser Funktion ist immer 1. Somit eignet sie sich eben perfekt für die Analyse von Stichproben, da die absolute Menge der Proben immer auf 1 normiert wird.

Ich möchte mich auf die beiden Formfaktoren und der Gleichung konzentrieren, die die allgemeine Normalverteilung auf jede spezielle Stichprobe anpasst. Im Folgenden werde ich Schritt für Schritt herleiten, wie man aus einer Menge von Messungen (zum Beispiel die Körpergröße von einhundert Frauen in Deutschland) die beiden Faktoren herleitet.

%matplotlib inline  
import matplotlib.pyplot as plt  
import numpy as np  
import math  
  
x = np.arange(-5, 5, 0.1)  
y = [1 / math.sqrt(2 \* math.pi) \* math.exp(-0.5 \* x\*\*2) for x in x]  
  
plt.plot(x,y, label="σ = 1 / µ = 0")  
plt.grid(True)  
plt.legend()  
plt.show()  
print("Abb 1: Normalverteilung für σ = 1 und µ = 0")

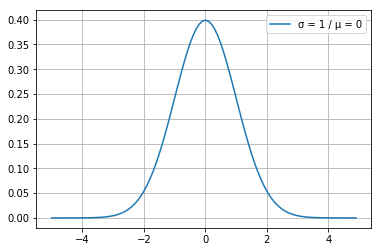


Abb 1: Normalverteilung für σ = 1 und µ = 0

## Formfaktor: der Erwartungswert

Der Erwartungswert ist schlichtweg der Mittelwert. Damit verschiebt man das Maximum der Kurve auf den Durchschnittswert.

%matplotlib inline  
import matplotlib.pyplot as plt  
import numpy as np  
import math  
  
x = np.arange(130, 210, 0.1)  
y\_klein = [1 / math.sqrt(2 \* math.pi) \* math.exp(-0.5 \* (x - 140)\*\*2) for x in x]  
y\_mittel = [1 / math.sqrt(2 \* math.pi) \* math.exp(-0.5 \* (x - 166)\*\*2) for x in x]  
y\_gross = [1 / math.sqrt(2 \* math.pi) \* math.exp(-0.5 \* (x - 180)\*\*2) for x in x]  
  
plt.plot(x,y\_klein, label="µ = 140cm")  
plt.plot(x,y\_mittel, label="µ = 166cm")  
plt.plot(x,y\_gross, label="µ = 180cm")  
plt.grid(True)  
plt.legend()  
plt.show()  
print("Abb 2: Normalverteilung für σ = 1 und µ = 140cm oder 166cm oder 180cm")

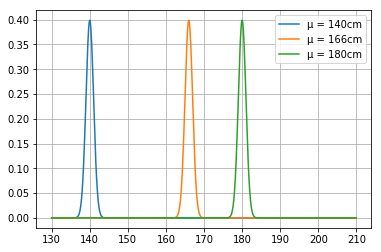


Abb 2: Normalverteilung für σ = 1 und µ = 140cm oder 166cm oder 180cm

## Formfaktor: die Varianz

Die Varianz gibt an, wie stark die Proben um den Mittelwert streuen. Wenn eine Schulklasse zum Beispiel eine Arbeit schreibt und 15 Kinder eine Eins und 15 Kinder eine Sechs schreiben, ist der Mittelwert eine 3, die Streuung ist aber extrem groß.

%matplotlib inline  
import matplotlib.pyplot as plt  
import numpy as np  
import math  
  
x = np.arange(-5, 5, 0.1)  
y\_klein = [1 / math.sqrt(2 \* math.pi \* 0.5\*\*2) \* math.exp(-0.5 \* (x)\*\*2 / 0.5\*\*2) for x in x]  
y\_mittel = [1 / math.sqrt(2 \* math.pi \* 1.0\*\*2) \* math.exp(-0.5 \* (x)\*\*2 / 1.0\*\*2) for x in x]  
y\_gross = [1 / math.sqrt(2 \* math.pi \* 2.0\*\*2) \* math.exp(-0.5 \* (x)\*\*2 / 2.0\*\*2) for x in x]  
  
plt.plot(x,y\_klein, label="σ = 0,5")  
plt.plot(x,y\_mittel, label="σ = 1.0")  
plt.plot(x,y\_gross, label="σ = 2.0")  
plt.grid(True)  
plt.legend()  
plt.show()  
print("Abb 3: Normalverteilung für µ = 0 und σ = 0,5 oder σ = 1.0 oder σ = 2,0")

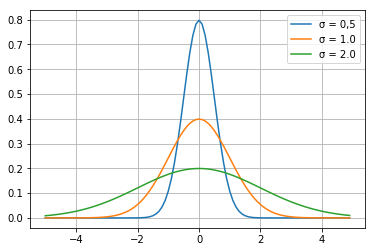


Abb 3: Normalverteilung für µ = 0 und σ = 0,5 oder σ = 1.0 oder σ = 2,0

## Ein Beispiel

Nehmen wir noch einmal die Durchschnittsgröße von Frauen in Deutschland. Laut www.laenderdaten.info ist sie 1,66m. Ich schätze nun ein von 12 und errechne die Wahrscheinlichkeit, dass ich als nächstes eine Frau mit 1,40m oder 1,66m oder 1,90m treffe.

%matplotlib inline  
import matplotlib.pyplot as plt  
import numpy as np  
import math  
  
sigma = 12  
mu = 166  
  
f = lambda x : 1 / math.sqrt(2 \* math.pi \* sigma\*\*2) \* math.exp(-0.5 \* (x - mu)\*\*2 / sigma\*\*2)  
  
x = np.arange(100, 220, 0.1)  
y = [f(x) for x in x]  
  
plt.plot(x,y, label="σ = 12 / µ = 166")  
  
h = 140  
plt.plot(h, f(h),'o-', label="{}cm = {:.2f}%".format(h, f(h)\*100))  
h = 166  
plt.plot(h, f(h),'o-', label="{}cm = {:.2f}%".format(h, f(h)\*100))  
h = 190  
plt.plot(h, f(h),'o-', label="{}cm = {:.2f}%".format(h, f(h)\*100))  
  
plt.grid(True)  
plt.legend(bbox\_to\_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)  
plt.show()  
print("Abb 4: Normalverteilte Körpergrößen von Frauen in Deutschland mit einem µ = 166cm und σ = 12")

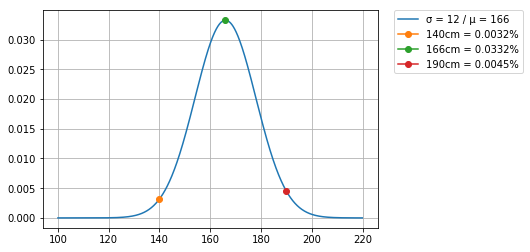


Abb 4: Normalverteilte Körpergrößen von Frauen in Deutschland mit einem µ = 166cm und σ = 12

Die Wahrscheinlichkeit, dass ich alle drei Frauen als nächstes treffe ist somit das Produkt aus den Einzelergebnissen:

## Herleitung der Formfaktoren

### Das Produkt

Nun stellt sich die Frage: Wie kann man aus einer Menge von Testergebnissen die beiden Formfaktoren und herleiten? Dazu gehe ich noch einmal zurück zum Beispiel, in dem man sehen kann, dass die Gesamtwahrscheinlichkeit das Produkt der Einzelwahrscheinlichkeiten ist:

### 

Wichtig ist hier, dass alle aus demselben Wahrscheinlichkeitsexperiment kommen müssen. Sie haben also alle das gleiche und das gleiche .

## Der Logarithmus

Man kann sich vergegenwärtigen, dass sich und gegenseitig nicht beeinflussen. Das verschiebt das Maximum der Kurve nach links oder rechts und das bestimmt die Höhe und die Breite der Kurve. Beide Parameter sind also voneinander entkoppelt und können einzeln betrachtet werden. Das gilt, solange alle aus dem gleichen Wahrscheinlichkeitsexperiment kommen und somit das gleiche und haben. Diese Vorüberlegung führt zum nächsten Schritt. Ich will wissen, wo das maximal wird. Dazu müsste ich die Funktion nach ableiten:

Das kann man machen ist wegen des Produkts der einzelnen aber nicht sehr lustig. Man kann hier aber zu einem kleinen Trick greifen. Wir suchen ja den Ort des Maximums und nicht den Betrag. Also können wir uns ja eine Funktion suchen, die ihr Maximum an der gleichen Stelle hat, die sich aber schöner ableiten lässt. Und die gibt es. Wir nehmen einfach von dem gesamten Ausdruck den Logarithmus:

### 

%matplotlib inline  
import matplotlib.pyplot as plt  
import numpy as np  
import math  
  
sigma = 2  
x = 144  
  
f = lambda mu : 1 / math.sqrt(2 \* math.pi \* sigma\*\*2) \* math.exp(-0.5 \* (x - mu)\*\*2 / sigma\*\*2)  
f\_ln = lambda mu : math.log(1 / math.sqrt(2 \* math.pi \* sigma\*\*2) \* math.exp(-0.5 \* (x - mu)\*\*2 / sigma\*\*2))  
  
mu = np.arange(100, 220, 0.1)  
  
plt.plot(mu, [f(mu\_i) for mu\_i in mu], label="W(µ,σ,x)")  
plt.plot(mu, [f\_ln(mu\_i) for mu\_i in mu], label="ln(W(µ,σ,x))")  
  
plt.yscale('symlog', linthreshy=0.01)  
plt.grid(True)  
plt.legend(bbox\_to\_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)  
plt.show()  
print("Abb 5: Normalverteilung und deren natürlicher Logarithmus für ein konstantes σ und x")

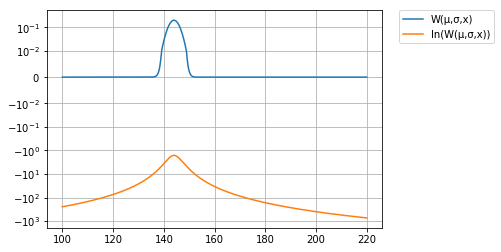


Abb 5: Normalverteilung und deren natürlicher Logarithmus für ein konstantes σ und x

Der Logarithmus hat die schöne Eigenschaft, dass er alle Rechenoperationen quasi um einen grad heruntersetzt. Somit können wir den Ausdruck oben in sehr einfache Teile zerlegen:

### 

### 

## Die Kettenregel

Mit der Kettenregel lässt sich der Ausdruck nun sehr einfach nach ableiten:

### 

### 

## Das Maximum

Das Maximum einer Funktion findet man bekanntlich, indem man sie einmal ableitet und zu null setzt:

### 

### 

### 

### 

## 

Die Formel für den Erwartungswert ist also der Mittelwert aller Einzelwahrscheinlichkeiten und das ist ja auch genau das Ergebnis, dass wir erwartet haben.

## Der nächste Parameter

Nun stellt sich die Frage, ob man den Logarithmus auch für die Herleitung der Varianz verwenden kann. Da wir oben aber festgestellt haben, dass die Parameter voneinander entkoppelt sind sollte auch das funktionieren.

%matplotlib inline  
import matplotlib.pyplot as plt  
import numpy as np  
import math  
  
mu = 166  
x = 144  
  
f = lambda sigma : 1 / math.sqrt(2 \* math.pi \* sigma\*\*2) \* math.exp(-0.5 \* (x - mu)\*\*2 / sigma\*\*2)  
f\_ln = lambda sigma : math.log(1 / math.sqrt(2 \* math.pi \* sigma\*\*2) \* math.exp(-0.5 \* (x - mu)\*\*2 / sigma\*\*2))  
  
sigma = np.arange(1.9, 250, 0.1)  
  
plt.plot(sigma, [f(sigma\_i) for sigma\_i in sigma], label="W(µ,σ,x)")  
plt.plot(sigma, [f\_ln(sigma\_i) for sigma\_i in sigma], label="ln(W(µ,σ,x))")  
  
plt.yscale('symlog', linthreshy=0.01)  
plt.grid(True)  
plt.legend(bbox\_to\_anchor=(1.05, 1), loc='upper left', borderaxespad=0.)  
plt.show()  
print("Abb 6: Normalverteilung und deren natürlicher Logarithmus für ein konstantes µ und x")

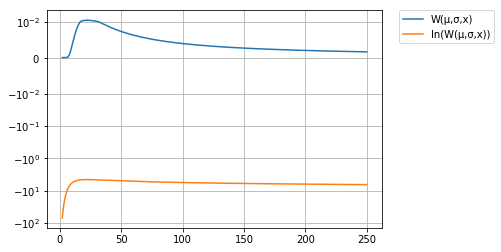


Abb 6: Normalverteilung und deren natürlicher Logarithmus für ein konstantes µ und x

Eine visuelle Inspektion ist ja bekanntlich kein Beweis, dass man auch hier den Logarithmus verwenden kann, um die Varianz herzuleiten. Somit glauben wir erst einmal der Behauptung, dass beide Kurven ihr Maximum an der gleichen Stelle haben und überprüfen, ob das Ergebnis plausibel erscheint:

## Simples Ableiten

### 

## Das Maximum

Auch diese Gleichung wird wieder zu null gesetzt:

### 

### 

Diese Formel kennt man als Statistiker auch. das ist nämlich die Standartabweichung. Also ein Maß für die Streuung von Stichproben um den Mittelwert. Damit sind die wichtigsten Werkzeuge für die Bewertung von Stichproben geschärft:

### Die Normalverteilung:

### 

### Der Erwartungswert:

### 

### Die Varianz:

### 