Mathematische und statistische Methoden für Pharmazeut*innen

Prof. Dr. Noemi Kurt
FB 12, Institut für Mathematik, Goethe-Universität Frankfurt

Sommersemester 2023

Vorlesung 11

Inhalt

- Häufigkeiten und Wahrscheinlichkeiten
- Berechnung von elementaren Wahrscheinlichkeiten
- Wichtige Wahrscheinlichkeitsverteilungen
- Normalverteilung, zentraler Grenzwertsatz

Lernziele

- Elementare Wahrscheinlichkeiten berechnen können
- Wichtige Wahrscheinlichkeitsverteilungen kennen
- Mit der Normalverteilung rechnen können
- Gründe für die Bedeutung der Normalverteilung kennen

Benötigte Vorkenntnisse

- Funktionen, Zahlenmengen, Mengenoperationen; Differential- und Integralrechnung

Relative Häufigkeiten und Wahrscheinlichkeiten

Intuition: Wenn die Zahl der Messungen n groß ist, so ist die relative Häufigkeit eines Messwertes $h_n(x)$ ungefähr gleich der theoretischen Wahrscheinlichkeit, dass dieser Messwert auftritt. Formal:

$$\lim_{n\to\infty} h_n(x) = \mathbb{P}(x),$$

bzw.

$$h_n(x) \approx \mathbb{P}(x).$$

Das $\mathbb{P}$ steht für Wahrscheinlichkeit (probability, probabilitas).

Dies nennt man auch (empirisches) Gesetz der großen Zahlen.

Häufigkeiten von Mengen/Klassen von Daten:

$$\mathbb{P}(A) \approx h_n(A)$$

welcher Anteil der Messwerte liegt in der Menge A, z.B. $A = [60, 70]$ wenn das Körpergewicht von Personen bestimmt wird.

Ereignisse und Mengen

Ein einzelner Messwert ist eine Zahl $x \in \mathbb{R}$, welche als Ergebnis eines zufälligen Vorgangs interpretiert wird.

Beispiel: Von n zufällig ausgewählten Probanden werden Alter x, Gewicht y und (systolischer) Blutdruck z gemessen.

Von Interesse sind Ereignisse, welche aus diesen Daten gebildet werden können. Dies sind Mengen:

- Der Proband ist mindestens 60: $A = \{x : x \geq 60\}$
- Der Proband wiegt unter 80 kg: $B = \{y : y < 80\}$
- Der systolische Blutdruck liegt zwischen 125 und 140: $C = \{z : z \in [125, 140]\} = \{z : 125 \leq z \leq 140\}$.

Davon abgeleitete Ereignisse:

- $A \cap B$: Ereignis A und B treten ein, also der Proband ist mindestens 60 Jahre alt und wiegt höchstens 80 kg.
- $A \cup B$: Ereignis A oder B (oder beide) treten ein
- A^c oder $\overline{A}$ das Gegenteil von A tritt ein
- $(A \cup B) \cap C^c \ldots$

Wahrscheinlichkeiten

Wahrscheinlichkeiten sind Zahlen zwischen 0 und 1. Manchmal werden sie auch in Prozent angegeben.

Für Ereignisse $A, B, C, A \cap B, \ldots$ schreiben wir

$$\mathbb{P}(A), \mathbb{P}(B), \mathbb{P}(C), \mathbb{P}(A \cap B) \ldots$$

für die entsprechenden Wahrscheinlichkeiten.

(Beispiele an der Tafel)

Wichtige Rechenregeln:

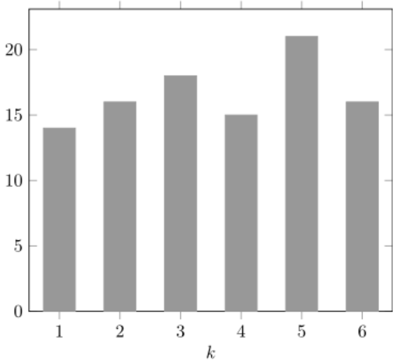
- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

Analoge Rechenregeln gelten für die relativen Häufigkeiten.

Verteilungen

Trägt man die experimentell beobachteten Häufigkeiten als Histogramm auf, erkennt man eine Häufigkeitsverteilung.

Beispiel: Häufigkeiten beim 100fachen Würfeln:

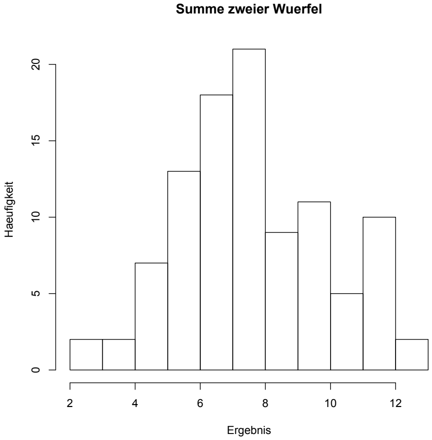


Alle Werte sind ungefähr gleich häufig: Gleichverteilung.

Verteilungen

Trägt man die experimentell beobachteten Häufigkeiten als Histogramm auf, erkennt man eine Häufigkeitsverteilung.

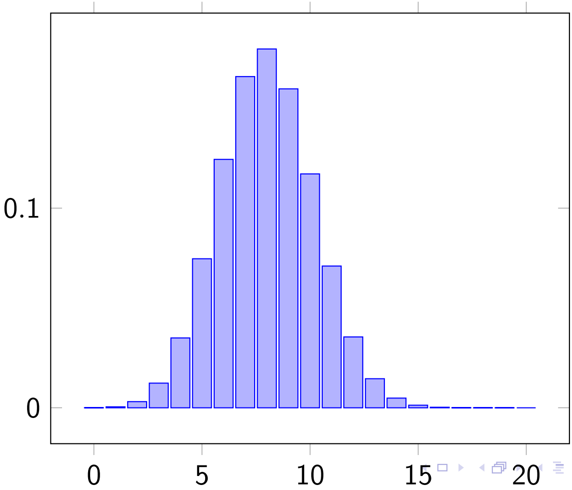
Beispiel: Häufigkeitsverteilung der Summe zweier Würfel:



Keine Gleichverteilung.

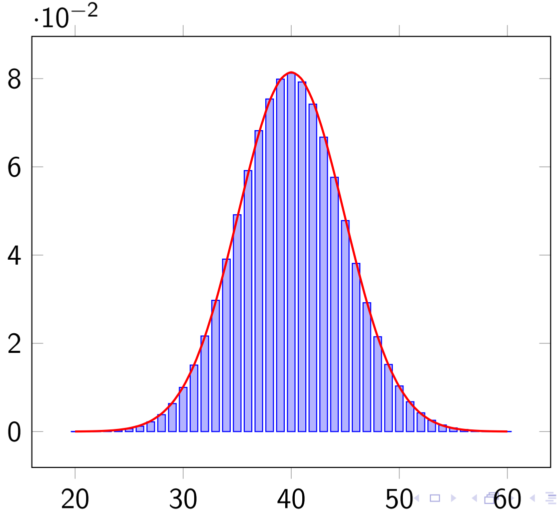
Verteilungen

Analog können Wahrscheinlichkeitsverteilungen dargestellt werden.
Beispiel: Theoretische Wahrscheinlichkeiten für k Behandlungserfolge, wenn 20 Personen behandelt wurden, und die Erfolgswahrscheinlichkeit 0.4 ist.



Normalverteilung

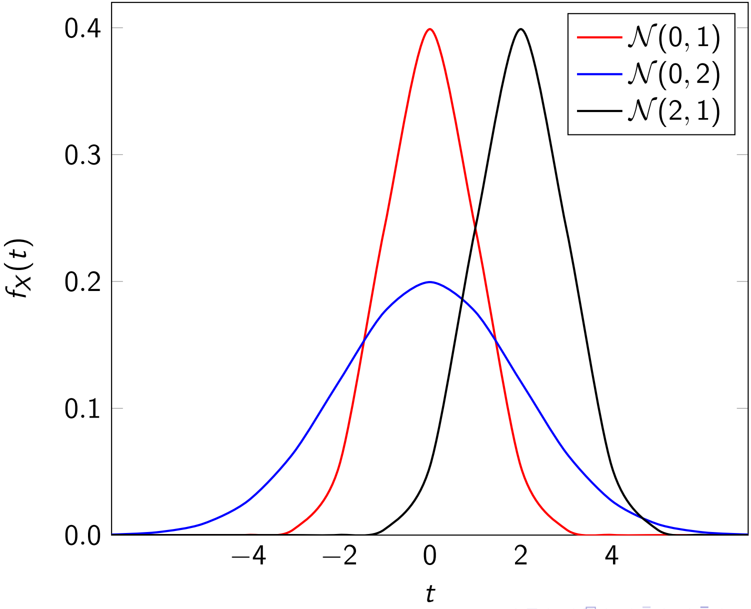
Hat die Häufigkeits- oder Wahrscheinlichkeitsverteilung (annähernd) die Form einer Gaußschen Glockenkurve, so handelt es sich (nährungsweise) um eine Normalverteilung.



Normalverteilung

- Die Stelle des Maximums (Werte mit der höchsten Wahrscheinlichkeit) liegt beim Mittelwert/Erwartungswert.
- Die Breite der Kurve wird durch die Varianz/Standardabweichung/Streuung bestimmt.
- Man sagt deshalb “die Daten folgen (näherungsweise) einer Normalverteilung mit Mittelwert μ und Varianz σ^2 (oder Standardabweichung σ).
- Faustregel: Folgen die Daten nähreungsweise einer Normalverteilung, so liegen 68.3% der Werte innerhalb einer Standardabweichung um den Mittelwert, also im Intervall $[\mu - \sigma, \mu + \sigma]$. Ca. 95.5% der Werte liegen innerhalb von zwei Standardabweichungen, also im Intervall $[\mu - 2\sigma, \mu + 2\sigma]$. (99.7 für 3σ.) Werte außerhalb dieser Intervalle sind sehr unwahrscheinlich.

Normalverteilung



Wann tritt eine Normalverteilung auf?

Grob gesprochen kann man immer dann annehmen, dass die Daten (nährungsweise) einer Normalverteilung folgen, wenn die gemessene Größe sich aus vielen kleinen Einzeleffekten zusammensetzt.

Formal gilt dann der Zentrale Grenzwertsatz.

Für praktische Zwecke sollte die Stichprobengöße (also die Zahl der Messwerte) mindestens $n = 30$ sein, damit die Normalverteilungsannahme vernünftig ist.

Rechnen mit der Normalverteilung

Leider ist das direkte Rechnen mit einer Normalverteilung schwierig. Theoretisch gilt:

$$\mathbb{P}(X \leq a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

wobei μ der Mittelwert und σ^2 die Varianz (σ die Standardabweichung) der normalverteilten Größe X ist.

Praktisch schaut man diese Wahrscheinlichkeiten in einer Tabelle nach, oder lässt sie mit dem Rechner berechnen.

Beispiel Excel: NORM.VERT(a; μ; σ; WAHR) gibt die Wahrscheinlichkeit aus, dass eine normalverteilte Größe X mit Mittelwert μ und Standardabweichung (!) σ kleiner oder gleich a ist.

Rechnen mit der Normalverteilung

Tabellen oder Programme wie Excel geben die Wahrscheinlichkeit

$$\mathbb{P}(X \leq a)$$

aus. Sind andere Wahrscheinlichkeiten gesucht, muss man die folgenden Rechenregeln verwenden:

- $\mathbb{P}(X \geq a) = 1 - \mathbb{P}(X \leq a)$
- $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a)$
- Beispiel an der Tafel

Standardisierung: Falls X normalverteilt mit Mittelwert μ und Varianz s^2 ist, so ist $Y = \frac{X-\mu}{s}$ normalverteilt mit Mittelwert 0 und Varianz 1, man spricht von Standardnormalverteilung.

Normalverteilung: Tabelle

$\Phi_{0,1}(x) = \mathbb{P}(X \leq x)$ für eine standardnormalverteilte Zufallsvariable X

x	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964

Zentraler Grenzwertsatz

- Der zentrale Grenzwertsatz besagt, dass für unabhängige Messungen der gleichen Messgröße

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma}$$

  ungefähr (für große n) standardnormalverteilt ist.
- Der zentrale Grenzwertsatz gilt universell, also egal welche Verteilung die x_i haben (solange sie unabhängig und identisch verteilt mit endlicher Varianz sind).
- Messwerte, welcher als Summe von unabhängigen, identisch verteilten Effekten interpretiert werden kann, ist (nach Reskalierung) ungefähr normalverteilt
- Wann immer viele unabhängige Ergebnisse aufsummiert werden, so ist das Ergebnis nach Reskalierung ungefähr normalverteilt.