


Umgang mit Zahlen – eine kleine Einführung in die Statistik

Einleitung

- 1  Der bisherige Physikunterricht mag den Eindruck vermittelt haben, dass alle Zusammenhänge genau bestimmt und berechenbar sind. Tatsächlich ist es aber so, dass man wegen Unbestimmtheiten unterschiedlicher Art mit Wahrscheinlichkeiten und Zufälligkeiten umgehen muss. Bisher haben wir uns in den Naturwissenschaften ausser bei Betrachtungen zur Messgenauigkeit noch kaum um diese Aspekte gekümmert, weil sonstige Zufälligkeiten durch die Wahl der Probleme und Systeme weitgehend ausgeschlossen wurden. Das Einführungsbeispiel dieses IF-Kurses (Osmose, Diffusion) zeigt jedoch, dass wir zwingend mit Wahrscheinlichkeiten umgehen müssen, sobald Prozesse mit vielen Körpern (Atome, Moleküle) untersucht werden. Je komplexer die Systeme werden, desto unausweichlicher wird es, die Zusammenhänge durch Wahrscheinlichkeiten auszudrücken; und fast jede wissenschaftliche Fragestellung ist ein komplexes Problem. Aus diesem Grund bedeutet wissenschaftliches Arbeiten immer auch korrekter Umgang mit Zahlen. **Biologische** Gesetze, **medizinische** Untersuchungen, **soziologische** und **psychologische** Studien, **ökonomische** und **ökologische** Sachverhalte, **Qualitätsanforderungen** und **Qualitätskontrollen**, Meinungsumfragen, usw. ... werden u.a. mit den Methoden der Statistik untersucht. Statistikvorlesungen bilden deshalb in vielen Studienrichtungen einen Teil der Grundausbildung.

Wir erachten es als sinnvoll, im Rahmen des Integrationsfaches auf diese universell eingesetzten Methoden hinzuweisen, so einen ersten Eindruck zu vermitteln.

Diese Einführung in die Statistik hat zwei Teile: Zuerst wird die **Verteilungen** einer **Messreihe** beschrieben. Sie werden lernen, was man unter den Stichworten **Mittelwert**, **absolute** und **relative Häufigkeit**, **Säulendiagramm**, **Normalverteilung**, **Varianz**, **Standardabweichung**, **Vertrauensintervall**, **Messfehler** versteht.

Im zweiten Teil werden zusätzliche Hilfsmittel vorgestellt, um den Zusammenhang **zweier Messgrößen** (zweier Messreihen) zu untersuchen, resp zu beschreiben. Wir sprechen vom **Korrelationskoeffizienten** und von der **Regressionsgerade** im **Punktdiagramm**.

Interview-Ausschnitt mit dem Statistiker *Jürg Hasler* im *Bund* vom 6. Juli 2002:

... Mediziner sind aufs Heilen spezialisiert und nicht aufs Rechnen. Ist Statistik für sie eine lästige Pflichtübung?

Die Mediziner wollen und müssen publizieren, da gehört Statistik meistens dazu. Ich meine, dass die Sensibilisierung für statistische Zusammenhänge in den letzten Jahren zugenommen hat. Die Kurse, die ich für Mediziner gebe, sind jedenfalls ständig ausgebucht.

Mit Statistik lässt sich grundsätzlich nicht beweisen, dass eine neue Therapie besser ist als die Standardtherapie. Man kann bestenfalls sagen, mit welcher Wahrscheinlichkeit der neue Ansatz dem alten überlegen ist.

Tatsächlich ist der statistische Nachweis ein Indizienbeweis. Er erlaubt nur bis zu einem gewissen Grad, die Spreu vom Weizen zu trennen. Umso wichtiger ist es, sich bei statistischen Tests nicht nur auf Irrtumswahrscheinlichkeit und die so genannten p-Werte zu konzentrieren. Man sollte auch die Vertrauensintervalle berücksichtigen und damit versuchen, die Wirksamkeit und Relevanz neuer Verfahren abzuschätzen.

Und warum begreifen nicht nur Mediziner solche Zusammenhänge schwer?

Die Fähigkeit, mit Wahrscheinlichkeiten und Unsicherheiten umgehen zu können, entsteht erst spät in der individuellen Entwicklung des Menschen. Sie zu erlangen, verlangt eine intensive Beschäftigung mit statistischen Konzepten und ihren Anwendungen in der praktischen Arbeit. In meinen Augen ist das eine interessante und wichtige Aufgabe. ...

Der Mittelwert einer Messgröße

- 2 \mathcal{L} Nehmen wir an, wir haben eine **Messreihe mit n Messwerten x_i** . Zuerst interessieren wir uns natürlich für den **Mittelwert**.

$$(1) \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- 3 ? Weil SchülerInnen sehr geübt sind im Berechnen von Notendurchschnitten, liegt die Wahl des ersten Beispielles (Kasten) auf der Hand. Wir können das ganze Prüfen-und-Notensetzen-Prozedere als Messvorgang verstehen, worin eine Note als Einzelmessung angesehen wird. Berechnen Sie den Notendurchschnitt aus den Angaben in der Tabelle.

Aus der Schule geplaudertes **Beispiel**: Die Tabelle zeigt die Zeugnisnoten, welche Wey im Fach Physik zwischen 1992 und 2001 in jenen Klassen setzte, die noch nach alter Maturitätsverordnung unterrichtet wurden.

Note	3	3.5	4	4.5	5	5.5	6
Anz.	28	171	542	625	433	126	19

Mittelwert =

Das Säulendiagramm

- 4 \mathcal{L} Der Mittelwert alleine sagt aber noch nicht alles aus. Sehr oft will man wissen wie die einzelnen Messwerte verteilt sind. Den ersten Eindruck gewinnt man durch eine Darstellung der Messwerte in einem **Säulendiagramm (Balkendiagramm, Histogramm)**.

Jeder Balken zeigt an, wie häufig ein Wert gemessen wurde. Oft werden dabei die Messwerte in **Klassen** zusammengefasst, die einen gewissen Messbereich abdecken. Die Häufigkeiten kann man als **absolute Häufigkeiten** (Anzahl) oder als **relative Häufigkeiten** (z.B. prozentuale Anteile) angeben.

Note	3	3.5	4	4.5	5	5.5	6
abs. H.	28	171	542	625	433	126	19
rel. H.	.0144	.0880	.2788	.3215	.2227	.0648	.0098
rel. H. in %	1.44	8.80	27.88	32.15	22.27	6.48	0.98

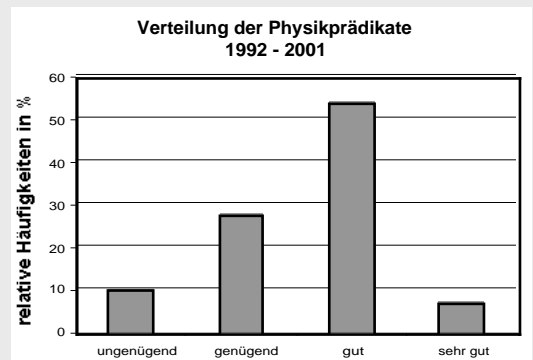
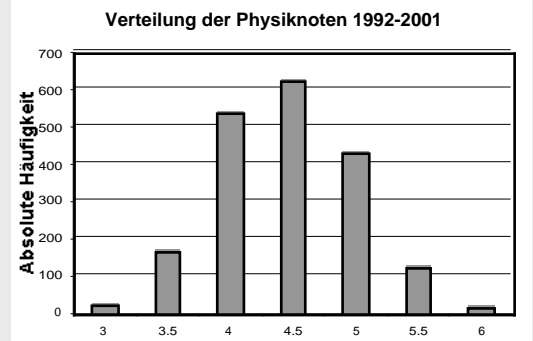
Normalverteilung, Varianz, Standardabweichung

- 5 \mathcal{L} Wenn die Unterschiede in den Messwerten zufällig sind und wenn die Messwerte in genügend grosser Zahl vorliegen, so ähnelt ein solches Balkendiagramm stets der „Glockenkurve“ der **Normalverteilung**, welche die Wahrscheinlichkeitsverteilung einer Zufallsgrösse bei „unendlich vielen“ Messungen angibt.

Eine Verteilungskurve kann schmal oder breit sein. Je schmaler die Verteilungskurve ist, desto genauer lässt sich das Ergebnis für eine neue Messung voraussagen. Man könnte also die Gewissheit, resp. die Unsicherheit eines Messwertes angeben, in dem man mit einer geeigneten Zahl die Breite der Verteilungskurve, resp. die **Streuung** der Messwerte, charakterisiert. Ein Mass für dies Streuung bekommen wir, indem wir den Mittelwert der quadrierten Differenzen zum Mittelwert berechnen. Diese Grösse heisst **Varianz σ^2** :

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


(Dass der Faktor 1/n nicht immer ganz korrekt ist, kummert Sie erst in der Statistikvorlesung an der Uni.)



Beide Säulendiagramme zeigen dieselbe Statistik, aber mit unterschiedlicher Angabe der Häufigkeiten und verschiedener Klasseneinteilung (Rubriken). Die Klasse *ungenügend* umfasst die Noten 3 und 3.5; *genügend*: 4; *gut*: 4.5 und 5; *sehr gut*: 5.5 und 6.

Die Quadratwurzel aus der Varianz nennen wir **Standardabweichung** σ . Sie gibt die Breite der Normalverteilungskurve an auf der Höhe der Wendepunkte.

Besteht eine Messreihe aus mindestens 30 Messwerten, welche zufällig streuen, so geben die Varianz und die Standardabweichung meist auch schon bei einer kleinen Zahl von Klassen eine sinnvolle Aussage.

- 6  Die Varianz kann auf unterschiedliche Weise berechnet werden. Wir können aus der Formel (2) eine weitere herleiten:

$$\begin{aligned}\sigma^2 &\approx \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2\bar{x} \cdot x_i + \bar{x}^2) \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \frac{1}{n} \sum_{i=1}^n x_i + \bar{x}^2 \frac{1}{n} \sum_{i=1}^n 1 \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \cdot \bar{x} + \bar{x}^2 \frac{1}{n} \cdot n\end{aligned}$$

$$(3) \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Das Vertrauensintervall

- 7  Für die Standardabweichung gilt: Das Intervall

$$(4) \quad [\bar{x} - \sigma, \bar{x} + \sigma]$$

ist das **Vertrauensintervall** für 68%-ige Wahrscheinlichkeit, dass ein weiterer Messwert innerhalb der Intervallgrenzen liegt.

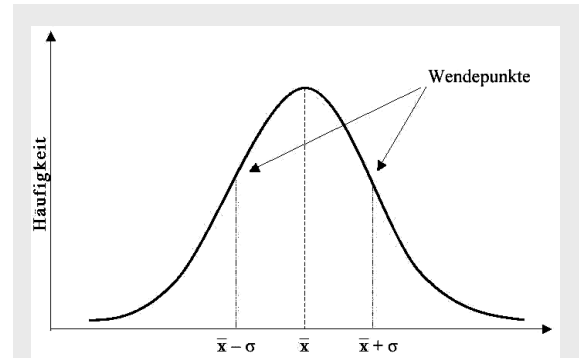
Ein Messwert liegt mit 95%-iger Wahrscheinlichkeit im Vertrauensintervall, wenn dieses wie folgt vergrößert wird:

$$(5) \quad [\bar{x} - 1.96 \cdot \sigma, \bar{x} + 1.96 \cdot \sigma]$$

- 8 ? Berechnen Sie im Notenbeispiel die Varianz und daraus auch die Standardbreite. Weil Sie den Mittelwert schon kennen, sind Sie mit der Formel (3) wohl schneller am Ziel als mit der Formel (2). Wie gross ist der prozentuale Anteil der Schüler mit einer Note im Intervall

$$[\text{Mittelwert} - \sigma, \text{Mittelwert} + \sigma] ?$$

Weil die Zahl der Klassen (d.h. die Zahl der möglichen Messwerte) klein ist, funktioniert die „68%-Regel“ noch nicht so gut.



Die Normalverteilung (Gauss'sche Glockenkurve) mit den charakteristischen Grössen Mittelwert und Standardabweichung.

Der Mittelwert und sein Messfehler

9 Versuchen Sie sich bei den folgenden Ausführungen nicht verwirren zu lassen. Mit der Standardweite können wir eine Aussage machen zur Wahrscheinlichkeit einer Einzelmessung. Macht man jedoch eine Messreihe mit n Messungen und fragt nach der Wahrscheinlichkeit, so ist zu erwarten, dass die Mittelwerte verschiedener Messreihen nicht mehr so stark streuen wie die Einzelmessungen. Die Statistiker beweisen, dass die **Standardabweichung der Mittelwerte** nur noch

$$(6) \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

beträgt. Da man als Messergebnis üblicherweise den Mittelwert einer Messreihe angibt, schreibt man das Messergebnis mit (6) als Messfehler auf.

$$(7) \quad \text{Messergebnis} = \bar{x} \pm \sigma_{\bar{x}}$$

10 Berechnen Sie in unserem Beispiel der Physiknoten die Standardabweichung des Notendurchschnittes. Geben Sie den Mittelwert mit vernünftiger Ziffernzahl und der Standardabweichung als Messfehler dar.

Zusammenfassung

Mittelwert $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

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

Standardabweichung
 $\sigma = \sqrt{\text{Varianz}}$

Vertrauensintervalle
 $[\bar{x} - c \cdot \sigma, \bar{x} + c \cdot \sigma]$

68%	c = 1
95%	c = 1.96
99%	c = 2.58

Messfehler des Mittelwertes
 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Verwendung von Taschenrechner und Excel

11 Mit der $\Sigma+$ **Taste der Taschenrechner** und mit **Excel** hat man gute Hilfen zur Berechnung statistischer Größen. Sie testen eines dieser Werkzeuge.

