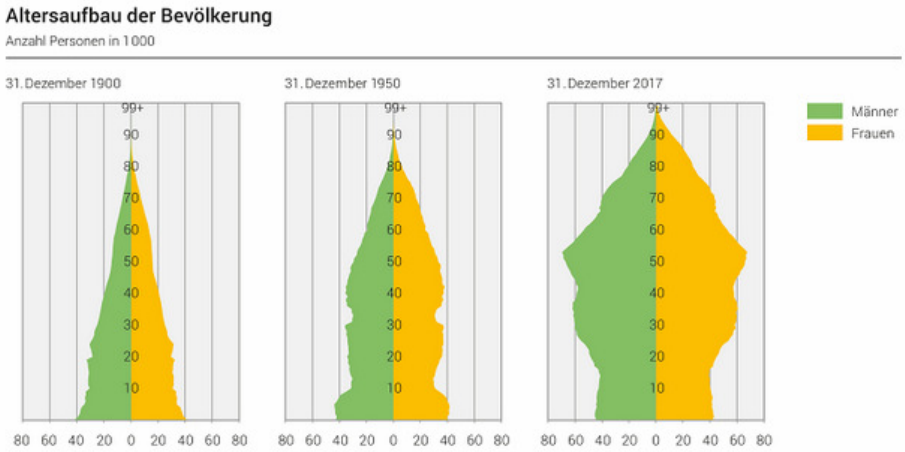
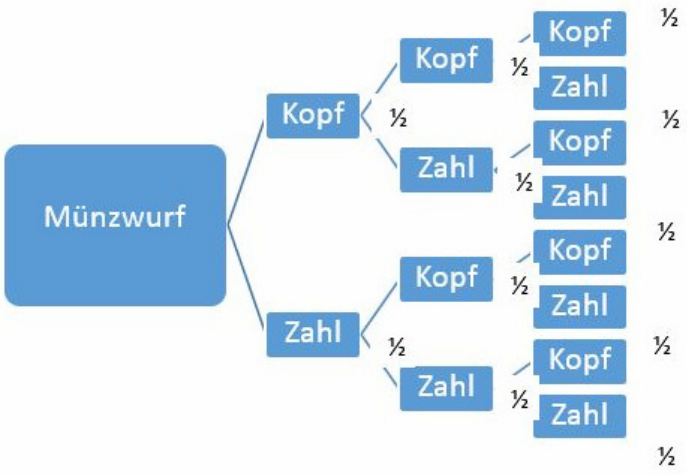
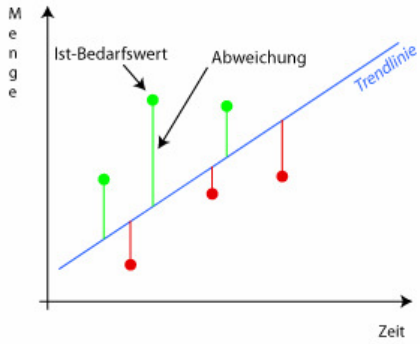
GIBB Modul 926, Teil 4

Stochastik







Statistik

&

Wahrscheinlichkeitrechnung

Definitionen

Stochastik kommt vom griechischen Wort στωχωσ (stochos) und bedeutet «klug vermuten, die Kunst des Vermutens». Damit werden Vorgänge untersucht, deren Ausgang nicht zum Vornherein bestimmt ist. Sie besteht aus

- **Statistik:**
  Der Begriff kommt vom lateinischen Wort status (Zustand). Die Anfänge der **Statistik** gehen auf die Volkszählungen vor unserer Zeitrechnung zurück.
  Erst om 18. Jahrhundert wurde sie zur selbständigen Wissenschaft, und ist eigentlich Teilgebiet der Mathematik; und
- **Wahrscheinlichkeitstheorie:**
  Sie untersucht Gesetzmässigkeiten zufälliger Ereignisse.
  Damit sollen auch Voraussagen für noch nicht aufgetretene Ereignisse getroffen werden können.

Zweck und Inhalt des Kapitels

- Grundbegriffe der Statistik kennen und anwenden
- Erhebung von ausgewählten Daten
- Aufbereitung
- Rückschlüsse auf die Gesamtheit ziehen und entsprechend handeln
- Daten orten, also zu betrachten, wo sie in der bestimmenden Umgebung angesammelt sind, wie sie sich dabei verteilen und was daraus an sinnvoller Information herausgelesen werden kann.

- Grundformel der Wahrscheinlichkeitsrechnung
- einfache Wahrscheinlichkeiten und deren Verknüpfungen betrachten.
- Berechnungsmöglichkeiten mit Werkzeugen wie dem Baumdiagramm oder mathematischen Formeln wie der Binomialverteilung und den Bernoulli - Experimente.

Auch für dieses Kapitel ist die Beherrschung der Bedienung Ihres Taschenrechners ein Muss.

Suchen Sie ein Manual und machen Sie sich ein Bild, ob der Rechner eingebaute Statistikfunktionen hat.

Beispiel 1 für Statistik: Datenerhebung von Körpergrössen

- Annahme: die Regierung hat beschlossen, für Lernende der GIBB Schul-Uniformen einzuführen.
- Sie haben den Auftrag, innert kurzer Zeit eine Abschätzung der Körpergrössen der Lernenden zu erstellen.

Statistische Begriffe:

Die etwa **7000 Berufsschüler** bilden die **Grundgesamtheit** oder **Population** Ihrer Umfrage.
Da Sie nicht alle befragen können, treffen Sie eine Auswahl. Sie beschränken sich aus Zeitgründen auf 2 Informatikerklassen pro Lehrjahr, das sind etwa 140 Lernende. In der Statistik nennt man **die Informatikerlernenden** die **Stichprobe** und die Zahl **140** ist den **Umfang der Stichprobe**.
Die Grösse (nicht etwa das Geschlecht oder die Korpulenz) stellt das interessierte **Merkmal** dar, und die **Ausprägung des Merkmals** ist die **Zahl in cm**. Das Ganze notieren Sie auf einem **Erhebungsformular**, die sogenannte **Urliste.**

Übung 1
Stellen Sie die obigen Begriffe in der Tabelle zusammen!

Wer oder was	**Statistischer Begriff**
Erhebungsformular, Fragebogen	
Alle Berufsschüler	oder
Die ausgewählte Klasse(n) Informatiker	
140 (Anzahl befragte Lernende)	
Körpergrösse	
z.B. 175 cm	

Vor jeder Datenerhebung muss klar sein, was damit bezweckt wird! Eine zweckmässige Struktur der **Urliste** kann viel nachträgliche Arbeit ersparen. Die Urliste kann die Form einer Strichliste haben. Sinnvoll ist es heute bestimmt, grössere Datenbestände elektronisch zu erfassen, ev. direkt in eine Datenbank aufzunehmen. Die Daten können einfacher verarbeitet werden, z.B. ordnen oder sortieren. In einer Umfrage können **mehrere** Merkmale vorkommen.

Sorgfältig muss ebenfalls die **Stichprobe** ausgewählt werden. Sie soll **repräsentativ** (stellvertretend ähnlich) für die Grundgesamtheit sein. Eine umfassende Stichprobe erfasst **zwischen 600 und 1200 Werte** (deshalb wird in öffentlichen Umfragen meist gesagt, dass z.B. 1150 Leute befragt wurden). Damit ist die Aussage in der Regel präzis genug, um ein Bild der Grundgesamtheit zu erhalten. Je weniger Werte erfasst werden, umso grösser ist die Gefahr, daneben zu liegen, also eine ganze Gruppe nicht berücksichtigt zu haben.

Aus Effizienz- und Kostengründen wird die **Grundgesamtheit** selten erfasst. Nur etwa bei Volkszählungen, die aber nicht jedes Jahr neu durchgeführt werden.
Sorgfältig ausgewählte und **gut vorbereitete Stichproben** sind besonders wichtig und sparen Aufwand, Kosten und Ressourcen.
Es ist einleuchtend, dass z.B. eine Autofirma nicht mit allen Neuwagen Crashtests durchführt...

Nach der Erfassunng der Daten ordnet man sie meistens, falls möglich oder sinnvoll.
Einfach zu ordnen sind zum Beispiel Zahlen, Datum, Ranglisten, Preise usw. Schwieriger wird hingegen die Bestimmung einer Ordnung bei Lieblingsfarben, Essensgewohnheiten, Reisezielen usw. Doch auch bei diesen Themen kann man bestimmte Reihenfolgen erstellen, je nachdem, was damit bezweckt wird.

Nehmen wir bei unserer Umfrage nach den Schuluniformen folgende Daten für **eine** Klasse an. Sie haben zusätzlich zur Körpergrösse noch das Merkmal Geschlecht (**m**/**w**) erfasst. 22 Personen:

178m	164w	176w	188m	173m	160w	176m	181m	177m	158w	169m
173m	178w	180m	173m	173w	179m	183m	168m	171w	173w	176m

Absolute und relative Häufigkeit

Die absolute Häufigkeit ist die Anzahl der Ereignisse, sog. Elementarereignisse.
Sie gibt an, wie viele Elemente mit jeweils dem **gleichen** Merkmal gezählt worden sind.
In unserem Beispiel der Schuluniformen kommt das Elementarereignis «Körpergrösse 158 cm» nur 1x vor. Die **absolute** Häufigkeit für dieses Ereignis ist also 1.

Die **absolute Häufigkeit** wird mit dem **Formelzeichen H_n** angegeben. Beispiel:

$H_n(A) = 1$ bedeutet: Die absolute Häufigkeit des Ereignisses A (Körpergrösse 158 cm) ist 1 (1x).

Die relative Häufigkeit ist der **Anteil** des Ereignisses A an der **Grundgesamtheit** oder an der Stichprobe. Die relative Häufigkeit wird mit dem **Formelzeichen h_n** angegeben.

$$h_n(A) = \frac{H_n(A)}{n}$$ n = Gesamtzahl aller Elemente in der Menge.

Bei unserem Beispiel ist die relative Häufigkeit des Ereignisses A (Körpergrösse 158 cm) folgende:

$$h_n(A) = \frac{H_n(A)}{n} = \frac{1}{22} = 0{,}04545 \; oder \; 4{,}545\,\%$$

Übung 2
Ordnen Sie die Daten nach Grösse und Geschlecht.
Erstellen Sie eine Strichliste (**linke** Hälfte der Tabelle auf nächster Seite).
Lassen Sie die rechte Tabelle noch leer.

Übung 3
Füllen Sie jetzt die **rechte** Tabelle der Übung 2 aus. Nehmen Sie die Geschlechter m und w nicht mehr auseinander. Die Häufigkeiten sollen sich nur auf die Körpergrösse beziehen.

Körpergrösse in cm	Stichliste: weiblich	Strichliste: männlich
158	I	
Befragte 22	Total w	Total m

Strichliste: Total m+w	Absolute Häufigkeit (Anzahl)	Relative Häufigkeit %
22	22	100 %

Sobald mehr als nur einige Daten erhoben sind, wird das Sortieren etwas mühsam und auch fehleranfällig. Der Gebrauch von Anfang an einer Software oder zumindest einer Strichliste ist sehr empfehlenswert. Nachdem unsere Daten erhoben und sortiert sind, können wir sie weiterverarbeiten.

Wie haben Sie Übung 2+3 empfunden?

☐ Spannend, möchte alle 140 Werte der Stichprobe sortieren ☐ Spannend, aber es genügt ☐ Ziemlich langweilig ☐ Muss ich nicht machen, das beherrsche ich blind

Wie beurteilen Sie die Gefahr, dass Fehler auftreten?

☐ Alles stimmte auf Anhieb ☐ Ich musste 2-3x etwas korrigieren ☐ Mein Ergebnis stimmt nicht ☐ Habe aufgegeben

Erfassung zufälliger Ereignisse: Würfel

Übung 4

Der «Zufallsgenerator“ eines Computers hat bei 5000 Versuchen 680-mal die Sechs, 1035-mal die Fünf, 910-mal die Vier, 705-mal die Drei, 770-mal die Zwei und 900-mal die Eins gewürfelt.
Berechnen Sie die relative Häufigkeit als Prozentsatz für alle Augenzahlen.
Falls Zeit: Machen Sie von Hand ein masstabsrichtiges Balkendiagramm für die 6 Augenzahlen.

Übung 5

Gedankenübung: 2 Würfel werden **gleichzeitig** geworfen.
Sogenannt **fairer oder idealer** Würfel: Jede Seite kommt exakt gleich wahrscheinlich.
Man betrachte die möglichen **Summen** der beiden Augenzahlen.
Tragen Sie jede mögliche Summe in die linke Tabelle ein, Häufigkeiten in die rechte:

Würfel 1 / Würfel 2	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

Summe	H_n	h_n (als Bruch)	h_n (in %)	= Wahrscheinlichkeit fürs Auftreten!
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
Total				

Beschreibende oder deskriptive Statistik

Daten übersichtlich darzustellen und Sachverhalte zu **beschreiben**, ist das Ziel der deskriptiven (beschreibenden) Statistik. Dazu dienen Tabellen, Kennzahlen und Grafiken.
Hier benötigen wir meist sog. **Lagewerte** und **Streuwerte:**

- **Lagewerte** geben an, **um welchen Wert** unsere Daten angesiedelt sind. Es sind oft **Mittelwerte**.
- **Streuwerte** geben an, wie die Daten um den Lagewert **verteilt** sind.

Lagewerte

Der wohl bekannteste und auch am meisten benützte Lagewert ist der

Mittelwert, Durchschnitt, das arithmetische Mittel

Definition: $\mu \; oder \; \bar{x}_A = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$

Übung 6

Berechnen Sie die den Mittelwert der Körpergrössen aller 22 erfassten Lernenden der Übung 2

$\bar{x}_A = \frac{x_1 + x_2 + x_3 + ... + x_{22}}{22} =$

Ihr Taschenrechner kann Ihnen ev. die Berechnung erleichtern: Statistikfunktionen!
Suchen Sie, ob und wie Sie auf Ihrem Taschenrechner Daten eingeben, speichern, auswerten können.

Tipps:

TI-30 eco	→	Wert eingeben, dann Taste Σ+ tippen, nächsten Wert eingeben, Σ+, usw. Nach Eingabe des letzten Werts zeigt das Display n = 22. Nach Eintippen von 2nd und $\bar{x}_A$ lesen Sie den Wert.
TI-30B/II usw.	→	Schalten Sie in den Stat-Modus und geben die Daten über DATA ein.
TI-30X Pro	→	data tippen, dann die Daten als L1(1) bis L1(22) eingeben. 2nd stat-reg/distr, dann auf 2:1-Var Stats, dann 4x ENTER
TI-Inspire	→	Über das Menu eine Liste erstellen, dann zur Wahrscheinlichkeitsrechnung

Auch für dieses Kapitel ist die Beherrschung der Bedienung Ihres Taschenrechners zwingend.

Suchen Sie ein Manual und machen Sie sich ein Bild, ob es eingebaute Statistikfunktionen gibt.

Übung 7

178	164	176	188	173	160	176	181	177	158	169
173	178	180	173	173	179	183	168	171	173	176

Tippen Sie die 22 Werte in Ihrem Rechner ein. Welchen Wert erhalten Sie für $\bar{x}_A$? . . . . . .

Was gibt die Berechnung von Σ_x an ? . . . . . .

Übung 8

Der «Zufallsgenerator für Würfel» eines Computers hat bei 400 Versuchen 71-mal die Sechs, 58-mal die Fünf, 84-mal die Vier, 70-mal die Drei, 65-mal die Zwei und 52-mal die Eins gewürfelt. Wie gross ist das arithmetische Mittel all dieser Würfe?

Hier kommt jeder Wert vielfach vor. Natürlich tippt man nicht alle 400 Werte einzeln ein, sondern multipliziert:

$$\bar{x}_A = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + \ldots + n_n \cdot x_n}{n_1 + n_2 + \ldots + n_n} =$$

Diese Art der Berechnung nennt man Ermittlung des «**gewichteten** arithmetischen Mittels».
Es wird berücksichtigt, **wie oft** (FRQ = Frequenz) dieselbe Grösse vorkommt.
Tipp:Falls vorhanden (z.B. TI-30), benützen Sie zur Eingabe die Hilfsfunktion (**2nd) FRQ** (Frequenz).

6	**2nd**	**FRQ**	**7**	**1**	**Σ+**	**5**	**2nd**	**FRQ**	**5**	**8**	**Σ+**	**…**
1	**2nd**	**FRQ**	**5**	**2**	**Σ+**		**2nd**	$\bar{x}$	**Anzeige**		**-> Resultat**	

Ihr Taschenrechner nimmt Ihnen die Berechnungen ab, Sie konzentrieren sich einzig aufs Eintippen!

Übung 9

Berechnen Sie das arithmetische Mittel aller möglichen Augensummen aus der Übung 5,
sowie die gefragten Werte, die Sie auf Funktionstasten des Rechners finden:

$\mu = \overline{x_A} =$ $\Sigma x =$ $\Sigma x^2 =$

Der Median z (oder Zentralwert)

Neben dem arithmetischen Mittel ist der Median der häufigste gebrauchte „Mittelwert“.

In einer **geordneten Reihe** von Stichproben (Elemente der Urliste) oder der Grundgesamtheit ist der Median der in der Mitte stehende Wert; links und rechts vom Median hat es gleich viele Elemente.

Ermittlung: Links und rechts werden immer gleichviele Puppen entfernt, bis nur noch eine Puppe (falls Anzahl ungerade) oder zwei Puppen (bei gerader Anzahl Puppen) verbleiben.

$$z = \frac{n+1}{2} \quad \Rightarrow \quad Bei\ 7\ Werte: \quad z = \frac{7+1}{2} = 4$$ Der Median ist der 4. Wert.

Bleiben bei einer **geraden** Anzahl Daten zwei Werte in der Mitte, so kann der Median bestimmt werden als

$$z = \frac{n+1}{2} \quad \Rightarrow \quad Bei\ 12\ Werten: \quad z = \frac{12+1}{2} = 6{,}5$$

- Einer der zwei Werte, das heisst der 6. oder der 7. Wert, oder
- Beide Werte, das heisst der 6. und der 7. Wert (zwei Mediane), oder
- Der arithmetische Mittelwert beider Mediane, falls dies sinnvoll ist.

Der Median besitzt gegenüber dem arithmetischen Mittelwert einige vorteilhafte Eigenschaften.

- Er kann auch dort eingesetzt werden, wo der arithmetische Mittelwert keinen Sinn ergibt, zum Beispiel, wenn man den Zahlenbereich verlässt und trotzdem eine Ordnung der Daten angeben kann.
- Er ist stabil gegen Ausreissern, das heisst gegen Daten, die sehr weit «daneben» sind.

Übung 10

Sie sollen in einer kleinen Umfrage bei einer Grossbank den durchschnittlichen Lohn der Angestellten bestimmen. Dazu fragen Sie 10 Personen nach ihrem Lohn und treffen dabei zufälligerweise auf den CEO. Hier die erhobenen Jahreslöhne in kFr:

48 114 153 87 68 145 129 39'000 117 98

Bestimmen Sie arithmetisches Mittel $\bar{x}_A$ und Median z.
Welcher Wert ist Ihrer Meinung nach repräsentativer für den Durchschnittslohn und warum?

Der Modus (Modalwert)

Der Modus ist der Wert in einer Stichprobe oder in der Grundgesamtheit, der **am Häufigsten vorkommt**. Als Lagewert hat auch er seine Wichtigkeit. Er kann zum Beispiel ziemlich weit entfernt von arithmetischem Mittel oder Median liegen und darf deswegen nicht vergessen oder übersprungen werden.

Allgemein werden bis zu zwei Modi akzeptiert. Die Verteilung heisst dann bimodal. Sind mehr als zwei vorhanden, so lässt man den Modus meistens unbeachtet.

Übung 11

Bestimmen Sie den Modus der Frauen, den Modus der Männer und den Modus beider Geschlechter zusammen aus folgenden Körpergrössen. Die Körpergrösse der Frauen stehen *kursiv und schattiert*.

158 *160* 164 *168* *169* 171 *173* *173* 173 *173* *173* *175* *176* 176
176 177 *178* 178 179 180 181 183 185 188 191

$Modus_w$ = $Modus_m$ = $Modus_{w+m}$ =

Wie bei allen Lagewerten muss immer gut überlegt werden, welcher Wert die Ortung unserer Daten am besten spiegelt. Das kann bei jeder Erhebung anders sein.

Wir haben somit einige Lagewerte betrachtet. Lagewerte geben an, um welchen Wert herum unsere Daten angeordnet sind.

Repetitionsfrage / Uebung

A) Sie werfen einen «normalen» Würfel 12-mal und erhalten folgende Augenzahlen:
2 5 4 5 3 6 3 1 5 4 6 1

Füllen Sie nachstehende Tabelle aus!

Geordnete Liste												

Augenzahlen		H_n	h_n
	1		
	2		
	3		
	4		
	5		
	6		
	Total		

Andere Werte	n	$\mu = \bar{x}_A$	z	Modus	Σx	Σx^2

B) In diesem Semester hat ein Lernender folgende Noten erhalten: 5,1 4,6 4,3
Eine vierte Probe ist noch ausstehend. Die Noten der Proben werden auf 1/10 genau erstellt, die Zeugnisnote auf ½ Note gerundet. 0,25 und 0,75 wird aufgerundet.

- Welche Zeugnisnote kann er noch bestenfalls erreichen?
- Welche Note muss der Lernende mindestens erreichen, wenn er eine Zeugnis-Note 5 erhalten will?
- Welche schlechteste Probe-Note darf er sich erlauben, wenn er seine bisherige Zeugnis-Note erhalten will?

C) Übung 12

Welcher Wert ist der **Median z** aus unserer Stichprobe? Umrahmen

158 160 164 168 169 171 173 173 173 173 173 175 176 176 176 177 178 178 179 180 181 183 185 188 191

Streuwerte

Streuwerte oder Streuungswerte geben an, wie die Daten um einen **Lage**wert herum (meist um den Mittelwert) **verteilt** oder gestreut sind.

Nehmen wir noch einmal die verschiedenen Werte der Stichprobe der Lernenden aus der Übung 2. Die Werte sind jetzt aufsteigend geordnet und um 3 Werte ergänzt.

158 160 164 168 169 171 173 173 173 173 173 175 176 176 176 177 178 178 179 180 181 183 185 188 191

Den ersten Streuungswert, den wir sofort erkennen können, ist die

Die Spannweite r

(r aus dem englischen range = Bereich).

Die Spannweite r gibt den **gesamten Bereich** an, innerhalb welchem die Daten angeordnet sind.

158 160 164 168 169 171 173 173 173 173 173 175 176 176 176 177 178 178 179 180 181 183 185 188 **191**

$r = x_{max} - x_{min}$ in unserem Beispiel: $r = 191 - 158 =$ **33**

Unser Beispiel ergibt eine sinnvolle Spannweite. Wir können uns sogar einfach vorstellen, wie alle Lernenden dieser Klasse im Grössenvergleich stehen würden und auch den Grössenunterschied von 33 cm zwischen dem Grössten und dem Kleinsten.

Doch manchmal ergibt die Spannweite nicht die gewünschte Übersicht. Denken Sie nur an die Löhne der Angestellten einer Grossbank aus Übung 8. Was sollen wir mit der Spannweite von
r = 39'000 – 48 = 38'052 kFr anfangen?

Wo liegt das Problem? Es liegt beim weit abgelegenen Lohn (=Ausreisser) des Herrn CEO. Auch die nur 48'000 Fr Lohn (eventueller Lohn des Reinigungspersonals, das nicht direkt von der Bank angestellt ist), dürften die Aussage über Banklöhne stark beeinflussen.

Wie bei der Übung 8 schon festgestellt, ist der **Median** stabil gegenüber Ausreissern. Der Median ist auch jener Lagewert, der mit dem Streuungswert „Spannweite“ am meisten zu tun hat.
Wir behalten also den stabilen Median und verknüpfen ihn so mit unseren Werten der Stichprobe, dass die Streuung sinnvoll wird.

Das Quartil (die Quartile)

Quartile sind weitere Kennwerte einer Stichprobe. Man teilt die Reihe der **geordneten** Werte in 4 Quartile ein.
Wichtig sind dabei das untere Quartil $Q_{0,25}$ und das obere Quartil $Q_{0,75}$.
In der Mitte liegt der **Median z**, manchmal auch als $\mathbf{Q_{0,5}}$ bezeichnet.
Unterhalb des $Q_{0,25}$ liegen damit 25 % der Werte und
oberhalb des $Q_{0,75}$ ebenfalls 25 % der Werte der Stichprobe.
Somit sind 50 % der Werte im Bereich $Q_{0,25}$ bis $Q_{0,75}$ vorhanden.
Dieser Bereich heisst Quartilabstand QA oder IQR = interquartile range.

Hat man genügend Werte bei einer Stichprobe erfasst, so kann man die Spannweite auf den Quartilabstand reduzieren und ist damit sicher, dass die **Ausreisser wegfallen**.

Die Quartile werden folgendermassen bestimmt:
Zuerst wird die Position des Quartils in der geordneten Reihe mit n Werten bestimmt.

$Q_{0,25} = \frac{n}{4}$ $\qquad$ $Q_{0,75} = \frac{n}{4} \cdot 3$ $\qquad$ Jeweils auf die nächsthöhere Zahl aufrunden.

Dann werden die **Werte** (nicht die Position) eingesetzt.

Beispiel

Die Lohnstichprobe der Übung 8 hatte folgende 10 Stichprobenwerte angegeben:

48	114	153	87	68	145	129	39'000	117	98	Liste ordnen!
48	68	**87**	98	114	117	129	**145**	153	39'000	geordnet

$$Q_{0,25} = \frac{n}{4} = \frac{10}{4} = 2,5 \quad gerun\ det \Rightarrow 3.\ Position \qquad Q_{0,25} = 87$$

$$Q_{0,75} = \frac{n}{4} \cdot 3 = 7,5 \quad gerun\ det \Rightarrow 8.\ Position \qquad Q_{0,75} = 145$$

$$Quartilabs\ ta\ nd\ QA = Q_{0,75} - Q_{0,25} = 145 - 87 = 58$$

Beispiel aus der Verkehrsplanung: Die Geschwindigkeit v_{85}

V steht für Geschwindigkeit. Wenn z.B. die Wirkung einer Tempo-30-Zone überprüft werden soll, so ist der sogenannte v_{85}-Wert wichtig als Überprüfungswert. Also nicht die Durchschnitts-geschwindigkeit, weil diese von einem einzelnen Raser verfälscht werden könnte.
Das heisst: v_{85} ist jene Geschwindigkeit, die von **85%** der Fahrzeuge **eingehalten** wird (während 15% sie überschreiten).

Übung 13 (müssen Sie nicht machen, ist zuwenig wichtig)

Bestimmen Sie die Quartile $Q_{0,25}$, $Q_{0,75}$ und den Quartilabstand der geordneten Reihe mit 25 Werten.

158 160 164 168 169 171 173 173 173 173 173 175 176 176 176 177 178 178 179 180 181 183 185 188 191

Übung 14 + 15 gestrichen

Wichtigster Streuwert: die Standardabweichung $S_{\bar{X}_A}$ *oder* σ_X

Wie aus dem Symbol ersichtlich, bezieht sich der Streuwert **Standardabweichung** auf den Lagewert des **Arithmetischen Mittels**. Definition:
Die Standardabweichung ist die Wurzel aus der mittleren quadratischen Abweichung der Daten von ihrem arithmetischen Mittel. Als Formel geschrieben heisst das:

$s_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	$\sum_{i=1}^{n}$	Bedeutet Summe, die sich auf jeden Wert (Index i von 1 bis n) bezieht.
	$(x_i - \bar{x})^2$	Quadrierte Differenz zwischen Wert x_i und arithmetischen Mittelwert $\bar{x}_A$
	n-1	Anzahl der Stichprobenwerte –1
Ausgeschrieben:	mit 3 bis n Werten:	$s_{\bar{x}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$

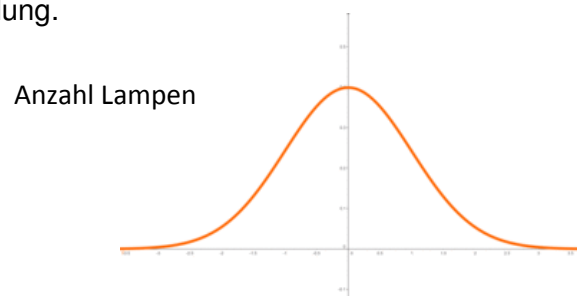
Oft wird auch nur durch n statt durch n-1 geteilt. Der Unterschied ist meist minimal, Details:

Summe der Werte	Σ_X
Anzahl Werte	n
Arithmetisches Mittel	$\bar{x}$ *oder* μ
Standardabweichung **der Stichprobe**	$S_{\bar{X}n-1}$ *oder* σ_{Xn-1}
Standardabweichung der **Grundgesamtheit**	$S_{\bar{X}n}$ *oder* σ_{Xn}

Übung 16
Berechnen Sie Mittelwert und Standardabweichung aus den 4 Noten **2, 6, 5, 3**.
Vergleichen Sie wenn möglich die Funtions des Taschenrechners mit einer Berechnung von Hand gemäss der Definitionsformel. Nötige Funktionen: Taste Σ+ für jeden Wert, dann Taste σ_{Xn-1}

Die Normalverteilung

Die Lebensdauer in Stunden einer Stichprobe von Glühlampen wird erfasst.
Sobald genügend Werte einer Stichprobe vorhanden sind, bemerkt man eine bestimmte Häufigkeits-Verteilung der Werte um das arithmetische Mittel herum, nach links und rechts abnehmend. Die Form der Verteilfunktion ist eine Kurve, die in der Mathematik als «**Gauss'sche Glockenkurve**» bekannt ist. Dies nenn man „Normalverteilung.



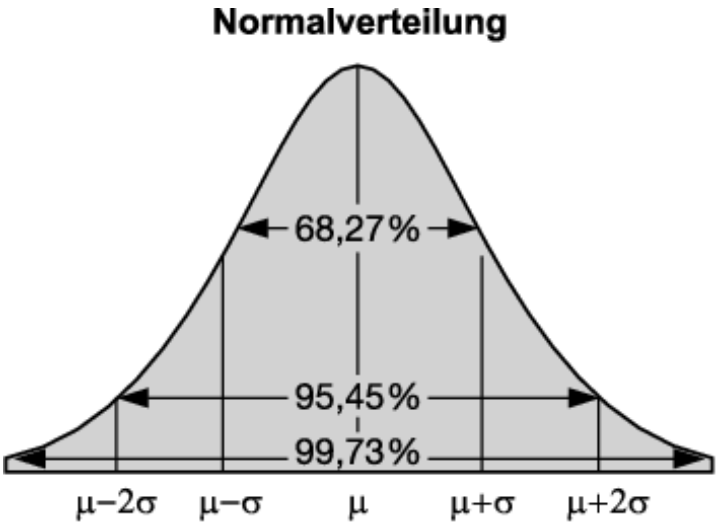
Eigenschaften dieser glockenförmigen Verteilung:

- Die Verteilung ist spiegelsymmetrisch um den Mittelwert μ
- Das Maximum der Verteilung liegt an der Stelle x = μ Mittelwert
- Die „Breite“ der Glocke wird durch die Standardabweichung σ bestimmt.

Die Standard-Normalverteilung

Sie ist das wichtigste Verteilmodell in der Statistik.
Obige Glockenkurve stimmt für die Brenndauer von Glühlampen. Würden wir die Körpergrösse unserer Lernenden darstellen, so erhielten wir ebenfalls eine glockenförmige Kurve, aber das arithmetische Mittel wäre natürlich nicht 1000 Stunden, sondern vielleicht 176 cm und die Standardabweichung z.B. 7 cm. Genauso wären die Werte für Batteriespannungen oder Toleranzen in der Uhrmacherkunst ganz andere, obschon die Kurvenform die gleiche bleiben würde. Eben die Normalverteilung.
Die **Standard-**Normalverteilung belässt die Glockenform und normiert die Achsen (macht sie unabhängig von den Einheiten).
Mit der Standardnormalverteilung lassen sich eine Vielzahl natur-, wirtschafts- und ingenieur-wissenschaftlicher Zusammenhänge, wie beispielsweise die Zufallsvariablen Körpergrösse, Gewicht, Messfehler, Zeiten bis hin zur Grösse von Sternen entweder exakt oder zumindest in guter Näherung beschreiben.
Nach der Normierung ist

- auf der x-Achse (Abszisse) die Anzahl der Standardabweichungen σ , die vom arithmetischen Mittelwert μ abweichen.
- Der Scheitelpunkt in der Mitte mit μ = 0 (Scheitelpunkt der Kurve)

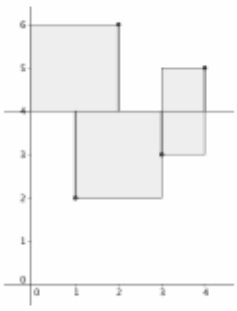


Da die Gauss’sche Glockenkurve symmetrisch ist, liegen 50 % der Werte links vom arithmetischen Mittelwert μ und 50 % sind rechts davon.

Bei einer guten Normalverteilung liegen 68 % der Werte innerhalb der einfachen Standardabweichung σ oder $S_{\overline{X}_A}$. Man spricht von der $^2/_3$ – Regel oder 1σ – Regel.

Aus einem andern Statistik-Kurs (M. Kriener Statistik):

Bei einer Liste von Zahlen $\{x_1, x_2, x_3, x_4\} = \{2, 6, 3, 5\}$ (das könnten zum Beispiel Ihre letzten Mathematiknoten sein) sind wir nicht nur am Mittelwert interessiert (der ist hier $\mu = 4$), sondern auch an der Streuung der Daten - hier hat man den Eindruck, dass die Noten ziemlich hin- und herspringen. Bei einem anderen Schüler sind die Noten (4.5, 3.5, 4, 4). Der gleiche Mittelwert, aber viel weniger Streuung. Wie können wir das Ausmass der Streuung mit einer Zahl erfassen? Einfach den Durchschnitt der **Abweichungen** $x_k - e\mu$ berechnen funktioniert nicht, denn $\Sigma(x_k - \mu) = 0$ (das ist die Robin-Hood-Gleichung). Wir möchten ausserdem, dass grosse Abweichungen stärker "zählen" sollen als kleine.



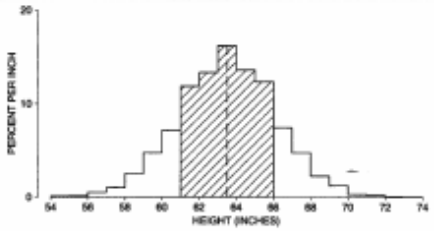
Wir haben gesehen, dass der Mittelwert μ die Summe der quadratischen Abweichungen $\sum(x_k - \mu)^2$ minimiert. Wir nennen den Mittelwert dieser Summe die **Varianz** und die Wurzel der Varianz die **Standardabweichung** einer Liste von Daten:

$$\sigma := \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2} \qquad (3)$$

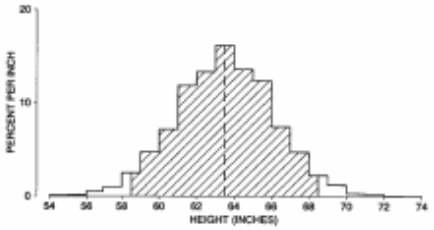
Beispiel: Für die Liste $(x_1, x_2, x_3, x_4) = (2, 6, 3, 5)$ von oben (Noten mit dem Mittelwert $\mu = 4$) ist die Standardabweichung

$$\sigma = \sqrt{\frac{2^2 + 2^2 + 1^2 + 1^2}{4}} = \sqrt{2.5} \approx 1.58$$

Was bedeutet diese Zahl? Ganz grob kann man sagen, dass die Standardabweichung "Normalität" misst - alles innerhalb *einer* Standardabweichung σ entfernt vom Durchschnitt ist noch "normal".



Beispiel: Die beiden Histogramme rechts zeigen zweimal die Grössenverteilung amerikanischer Frauen. Die gestrichelte Linie ist der Mittelwert, und der schattierte Bereich im oberen Histogramm ist der Bereich, der maximal eine Standardabweichung σ von μ entfernt ist - das sind etwa zwei Drittel aller Frauen. Der schattierte Bereich im unteren Histogramm ist der Bereich, der maximal zwei Standardabweichungen 2σ von μ entfernt ist - das sind schon 95% der Frauen.



Es gibt eine Daumenregel, die bei sehr vielen Daten gut funktioniert, nämlich bei **normalverteilten** Daten. Eigenschaften sind dann normalverteilt, wenn sie das Resultat von vielen voneinander unabhängigen Einflüssen sind. Körpergrösse und Intelligenz sind zum Beispiel normalverteilt. Die Daumenregel geht so:

- etwa 68% der Daten liegen innerhalb einer Standardabweichung vom Mittelwert - d.h. im Intervall $[\mu - \sigma; \mu + \sigma]$ liegen etwa 68% der gesamten Daten. Das entspricht der schattierten Fläche im oberen Histogramm.
- etwa 95% liegen innerhalb von zwei Standardabweichungen vom Mittelwert - d.h. im Intervall $[\mu - 2\sigma; \mu + 2\sigma]$ liegen etwa 95% der gesamten Daten. Das entspricht der schattierten Fläche im unteren Histogramm.
- etwa 99.7% einer Population liegen innerhalb von drei Standardabweichungen vom Mittelwert. Hier müsste man praktisch die gesamte Fläche schattieren.

Übung . Berechnen sie die Standardabweichung der folgenden Listen:

a) {4, 5, 4.2, 4.8}

b) {10, 8, 13, 11, 12, 9, 7}