6 Umgang mit Zahlen und Daten

6.1 Datensätze

■ (6.1.1) Die Ergebnisse von Beobachtungen und Messungen liegen in der Regel in Form von *Datensätzen* vor.

> Ein (numerischer) **Datensatz** ist so etwas wie eine zweispaltige Liste oder "Wertetabelle". Dabei ist es wichtig, zu unterscheiden, ob die Eintragungen der ersten Spalte (unabhängige Variable) eine inhaltliche Bedeutung haben - etwa Zeitpunkt der Temperaturmessung - oder ob sie reine Numerierungsfunktion für die Zahlwerte der zweiten Spalte haben! Wir sagen im ersten Fall Datensatz vom "Auszählungstyp" und im zweiten vom "Zeitreihentyp" .

(6.1.2) Wir betrachten hier vornehmlich den **Auszählungstyp**. Einen solchen pflegt man in der Regel wie folgt auf folgende Weise weiter zu verarbeiten:

♦ Zerlege den Wertebereich (die möglichen Werte) disjunkt in angemessene Teile.

◊ Im diskreten Fall kann das vorgegeben sein durch die überhaupt möglichen Werte, sonst ist künstlich zu diskretisieren.

♦ Die Teile werden geeignet benannt.

♦ **Zähle aus**, wieviele Daten in jeden dieser Teile fallen.

♦ Das gibt die "absoluten Häufigkeiten". Etwa N_i = Zahl der Daten, die in den i-ten Bereich fallen.

♦ Daraus kann man die "relativen Häufigkeiten" bilden: Ist N die Gesamtzahl aller Fälle, und N_i die absolute Häufigkeit, dann ist $n_i = \frac{N_i}{N}$ die "relative Häufigkeit" für den i-ten Bereich. (Oder: Die "Quote" dieses Falles)

♦ (6.1.3)Veranschaulicht ("graphisch dargestellt") werden beide Typen von Häufigkeiten in der Regel durch Histogramme (Stabdiagramme).

> ◊ Ein wichtiger Unterschied bei der Darstellung der beiden Häufigkeiten: Läßt man die Gesamtzahl N wachsen, dann wächst das Digramm für die absoluten Häufigkeiten mit, wogegen die Form der relativen Häufigkeiten sich stabilisieren sollte! Insbesondere kann man für die relativen Häufigkeiten N nach unendlich gehen lassen.
>
> □ Wir betrachten hier meist "zahlwertige Datensätze". Daneben gibt es "vektorwertige" und solche mit noch anderen Werten. Was ist damit gemeint? Statt "zahlwertig" sagt man meist "numerisch".

◊ (6.1.4) Datensätze, bei denen die Werte aus Zahlpaaren bestehen, veranschaulicht man gerne als "Punktschwarm in der Ebene". Diese Darstellung ist sowohl beim Typ Zeitreihe wie auch Auszählungstyp sinnvoll, sofern **-** wie gesagt - **die Werte als Zahlenpaar** vorliegen. Beispiel: Jedem Monat wird das Zahlpaar (mittlere Temperatur, mittlerer Niederschlag) zugeordnet.

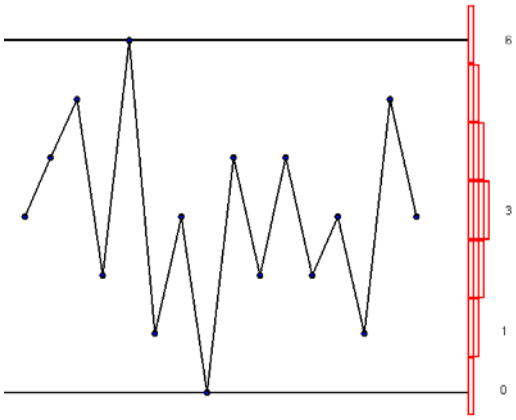
(6.1.5) Beispiel eines kleinen Datensatzes mit Zahlwerten von 0 bis 6:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
3	4	5	2	6	1	3	0	4	2	4	2	3	1	5	3

oder (selbsterklärend) kurz:

3	4	5	2	6	1	3	0	4	2	4	2	3	1	5	3

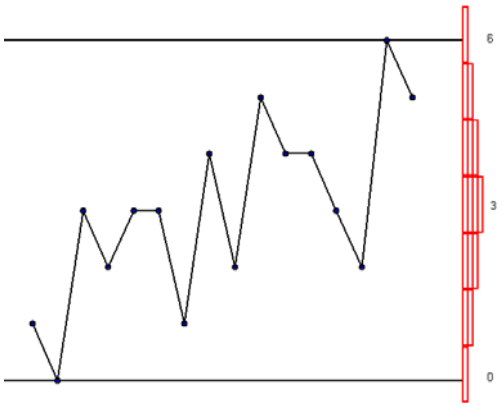
Also $a_1 = 3$, $a_2 = 4$ usw. Diesen tragen wir einmal als "Zeitreihe" auf, also horizontal die Hausnummer i und vertikal der Wert a_i. Rechts daneben dann die Auszählung der Werte als Histogramm. (Etwa $N_3 = 4$ gleich Zahl der Daten mit Wert 3): :



Jetzt dassselbe für einen zweiten Datensatz:

1	0	3	2	3	3	1	4	2	5	4	4	3	2	6	5

mit der graphischen Darstellung



Die Auszählung ergibt in beiden Fällen dasselbe. Aber im zweiten Fall hat der Index i, die Hausnummer, offenbar eine inhaltlich Bedeutung. Denn der Wert scheint im Mittel mit ihr anzusteigen. Wir haben es hier eher mit einem Datensatz vom Zeitreihentyp zu tun.

☐ Wie groß sind für beide Datensätze die absoluten und die relativen Häufigkeiten?

☐ Nehmen Sie eine zweite Auszählung für die Wertebereiche $0 \leq a_i \leq 2$, $3 \leq a_i \leq 4$ und $a_i \geq 5$ vor. Wie sehen dann die relativen Häufigkeiten (Quoten) aus?

6.2 Das Codierungsproblem der **beschreibenden Statistik**

■ (6.2.1) Wir betrachten einen numerischen Datensatz vom Auszählungstyp. Er enthält sehr viel Information, wobei in der Regel jedoch Unwichtiges mit wenig und schwer erkennbar Wichtigem vermischt ist.

(6.2.2) **Ziel:** Beschreibe und charakterisiere den Datensatz durch wenige (in der Regel 2) Zahlangaben, die möglichst viel Information liefern sollen:

Man nimmt dazu zwei Zahlen "Mittelwert und Streuung". Die erste Zahl beschreibt die Lage, den "Ort" der Daten auf der Zahlachse in Form eines Ersatzpunktes. Die zweite Zahl gibt ein Maß dafür, wie weit die Daten um diesen Ersatzpunkt herum streuen.

(6.2.3) Die **Berechnungsformeln** (der beiden Größen)I:

$a = (a_1, a_2, ..., a_N)$ numerischer Datensatz		
$\bar{a} = \mu(a) = \frac{1}{N}(a_1 + .. + a_i + ... + a_N) = \frac{1}{N}\sum_{i=1}^{N} a_i$		arithm. Mittel
$Var(a) = \frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2$	$\sigma(a) = \sqrt{Var(a)}$	Streuung und Varianz

Beachten Sie die Reihenfolge: a muss vorgegeben sein. Damit erhält man $\bar{a}$. Diese Zahl wird benötigt, um Var(a) zu berechnen. Und damit folgt σ.

☐ Vergleichen Sie diese Formel mit der vektoriellen Schwerpunktsformel und der Formel für ein gewichtetes Mittel.

☐ Zum Stichwort "Ersatzpunkt". Es gilt $N\bar{a} = \sum_{i=1}^{N} a_i.$, so dass man in Rechenausdrücken die Summe über alle Daten einfach durch Nā ersetzen kann. Welche Ersetzung erlaubt die Streuung $\sigma(a)$?

☐ **Wichtig**: Wie gewinnt man die beiden Größen mit dem Taschenrechner? Das sollte man können!

☐ Bestimmen Sie Mittelwert und Streuung der beiden Datensätze aus (6.1.5). Wieso müssen die Ergebnisse übereinstimmen?

☐ Angenommen Sie haben eine Auszählung des Datensatzes. Wann kann man diese benutzen, um Mittelwert und Streuung zu bestimmen? Und wie lauten dann die Formeln?

(6.2.4) Zur Verdeutlichung der Bedeutung der beiden Größen kann man den Datensatz wie folgt weiterverarbeiten:

Wir tragen alle Daten als Punkt auf der Zahlengeraden auf.

♦ **Dann wählen wir ā als neuen Ursprung oder Aufpunkt auf der Zahlengeraden und σ als neue Einheit.**

Jetzt können wir die Daten neu codieren, indem wir schreiben

$a_i = \bar{a} + \alpha_i \sigma$ mit eindeutig bestimmtem α_i.

Das gibt auf der Zahlgeraden die folgende Wegbeschreibung:

(1) Gehe (auf der Zahlgeraden) zum neuen Ursprung ā, und (2) von dort um α_i Einheiten σ weiter. Entscheidend ist: Die neuen Koordinaten α_i sind Zahlen in der Größenordnung von 1. Völlig unabhängig davon, wie groß die a_i selbst sind.

Oder: **Die α_i beschreiben die Abweichung vom Mittelwert in (typischerweisen kleinen) Vielfachen von σ**

(6.2.5) **Wichtig.** Überschlägig kann man sagen: Für etwa (50 ±20)% der Daten hat man $|\alpha_i| < 1$ und für etwa 90% hat man $|\alpha_1| < 2$.

Das ist die gewünschte Umcodierung des Datensatzes!

(6.2.6) Explizit wird man die α_i meist nicht ausrechnen. Sie zeigen zunächst nur die Bedeutung von σ auf. Graphisch zeichnet man meist von ā ausgehend die Änderung $\pm\sigma$ ein. Vgl. die Figur aus (6.2.8).

Ausnahme: Wenn man zwei Datensätze vergleichen will, deren Zahlenwerte und Streuung sehr unterschiedlich ist. Dann ist solche Normierung angebracht.

☐ Sie haben zwei Datensätze. Einer gibt die Längen einer großen Zahl von Elefanten wieder, der andere die von Mäusen. Sie möchten beide vergleichen, etwa untersuchen, ob die jeweiligen Längenunterschiede "dieselbe Form" besitzen. Wie werden Sie vorgehen?

☐ (6.2.7) Die Zahlen α_i bilden selbst wieder einen numerischen Datensatz. Was erwarten Sie für dessen Mittelwert und Streuung? Versuchen Sie das zu beweisen.

(6.2.8) Beispiel: Der Datensatz

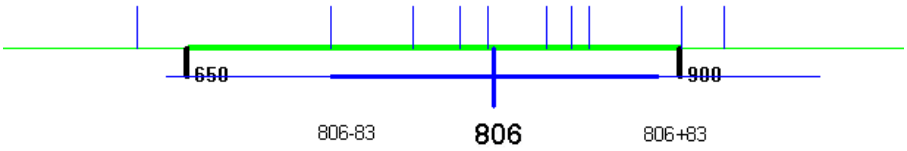
$$923 \quad 789 \quad 845 \quad 803 \quad 901 \quad 723 \quad 854 \quad 765 \quad 625 \quad 833$$

hat Mittelwert und Streuung (Datenstreuung):

$$\boxed{806 \pm 83}$$

D.h. die Daten liegen typischerweise zwischen 720 und 890.

In der nachfolgenden Figur ist der Bereich der Zahlengeraden zwischen 650 und 900 samt einer kleinen Umgebung grün aufgetragen. Als blaue Striche die 10 Daten des Datensatzes.



Unterhalb der Zahlengeraden die Lage des Mittelwertes und der 1-σ-Streubereich. Man erkennt den einen Ausreißer bei 625, der die Streuung vergrößert.

Jetzt der Datensatz der "Abweichungen vom Mittelwert" (etwa 923-806=117):

$$117 \quad -17 \quad 39 \quad -3 \quad 95 \quad -83 \quad 48 \quad -41 \quad -181 \quad 27$$

und normiert auf die Streuung:

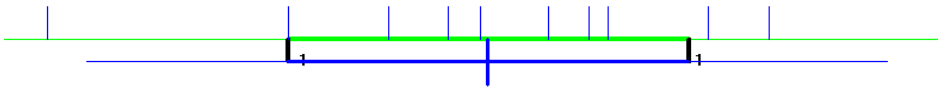
$$\frac{117}{83} \quad \frac{-17}{83} \quad \frac{39}{83} \quad \frac{-3}{83} \quad \frac{95}{83} \quad \frac{-83}{83} \quad \frac{48}{83} \quad \frac{-41}{83} \quad \frac{-181}{83} \quad \frac{27}{83}$$

Ausgerechnet gibt das den Datensatz der α_i. (Etwa $117 \approx 806 + 1.4 \cdot \sigma$)

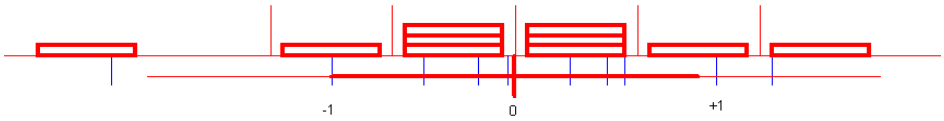
$$1.4 \quad -0.2 \quad 0.5 \quad -0.04 \quad 1.1 \quad -1.0 \quad 0.6 \quad -0.5 \quad -2.2 \quad 0.3$$

Das nächste Bild zeigt den Datensatz dieser Koordinaten im Bereich um α =-1 und +1.

Dieser neue Datensatz hat Mittelwert 0 und Streuung 1 (Winzige Abweichungen davon sind durch Rundungsfehler verursacht!)



Das nächste Bild zeigt noch (selbsterklärend) eine Auszählung des Datensatzes der α_i.



☐ Wie sehen hier die in (6.2.1) eingeführten zerlegenden "Teilbereiche" aus?

6.3 Wahrer Wert, Schätzwert und Fehler

■(6.3.1) Bei der Verarbeitung von Messergebnissen hat man es in der Regel mit folgender Situation zu tun:

> Man mißt eine Größe, von der man animmt, dass sie nach ausreichender Idealisierung einen "wahren Wert" besitzt. Das sei a_{Wahr}. Aber jede Messung ist ungenau. Man macht einen Fehler. Das Messergebnis der i-ten Messung sei a_i. Dann haben wir $\boxed{a_i = a_{Wahr} + f_i}$, wenn f_i der Fehler ist. Man hofft, dass man irgendwie zu a_{Wahr} gelangt, sofern man nur häufig genug misst.

Das ist vielfach ein schwieriges Problem, dass sorgfältiger fallspezifischer Analyse des zugehörigen "systematischen Fehlers" bedarf. Ein solcher kann etwa durch falsche Eichung eines Messgerätes entstehen. Daneben gibt es "statistische Fehler", die durch nicht kontrollierbare Ungenauigkeiten beim Messvorgang hervorgerufen werden. Nur mit ihnen wollen wir uns befassen.

Wie streuen die Daten (eines Datensatzes a von Messwerten) **um den wahren Wert?** Das ist eine analoge Frage zur Streuung der Daten um das Datenmittel. Analog zu Datensatzstreuung wird man die Streuung der Daten um den wahren Wert durch folgende Formel beschreiben:

$$\sigma_{Wahr} = \sqrt{Var_{Wahr}} \quad \text{mit} \quad \boxed{\mathrm{Var}_{Wahr}=\frac{1}{N}\sum_{i=1}^{N}(a_i\text{-}a_{Wahr})^2}$$

D.h. $\bar{a}$ ist durch a_{Wahr} ersetzt worden. Aber es besteht normalerweise ein großer Unterschied zwischen den beiden Größen:

> Bei gegebenem Datensatz kann man $\bar{a}$ berechnen und damit auch $\sigma = \sigma_{Daten}$, die Abweichung der Einzeldaten vom Mittelwert. Dagegen kann man a_{Wahr} **nicht mit Hilfe des Datensatzes bestimmen,** höchstens schätzen. Und.dann kann man σ_{Wahr} - die Streuung der Einzeldaten um den wahren Wert - auch nicht mit Hilfe der gegebenen Formel berechnen.

Aber kann man mit Hilfe des Datensatzes alleine vielleicht doch eine Schätzung dieses Fehler bestimmen, also eine Schätzung von σ^S_{Wahr} von σ_{Wahr}?

(6.3.2) Sofern kein systematischer Fehler vorliegt, stehen wir vor dem folgenden zentralen Problem:

> Gegeben ein Datensatz $(a_1, ..., a_N)$ von Messergebnissen der gesuchten Größe. Was kann man dann **nur** mit Hilfe dieser Daten über a_{wahr} herausbekommen? Und wie genau wird die Voraussage sein? .

Dazu kann und sollte man wie folgt vorgehen. (Auf einen Beweis oder eine genauere Rechtfertigung verzichten wir hier).

- ♦ Man habe einen Datensatz von N Messungen a_i der gesuchten Größe.
- ♦ Bilde das Datenmittel $\bar{a}$ des Datensatzes. Das ist ein **Schätzwert** für a_{Wahr}.
- ♦ Jetzt braucht man noch einen Schätzwert für die Größe der Fehler der Einzeldaten. Das ist **fast** die (oben besprochene) Datensatzstreuung. Diese selbst sollte man nicht nehmen, sondern stattdessen ("1/(N-1)-Regel! S für *Schätzung*)

$Var_{N-1}(a) = \frac{1}{N-1}\sum_{i=1}^{N}(a_i - \bar{a})^2 = ...$ $\quad \sigma_{N-1}(a) = \sqrt{Var_{N-1}(a)}$	**Datenschätzung** der Streuung $\sigma^S_{Wahr} \approx \sigma_{N-1}(a)$

N muss daher mindestens 2 sein! Beide Schätzungen benötigen nur den Datensatz selbst!

(6.3.3) Erneut hat man damit eine Codierung der Messergebnisse der folgenden Art:

$\boxed{a_i = \bar{a} + \varepsilon_i \sigma^S_{Wahr}(a)}$

a_i Das i-te Messergebnis
$\bar{a}$ bekannt, Schätzwert des wahren Wertes
$\sigma_{N-1} = \sigma^S_{Wahr}(a)$ bekannt
ε_i unbekannt, **aber** von der Größenordnung 1

(6.3.4) Und wie steht es mit unserem Wissen über die Genauigkeit von a_{wahr}? Hierzu geben wir ein **wichtiges Resultat** erneut ohne Beweis.

Resultat zur "Streuung der Mittelwerte" oder "$1/\sqrt{N}-$Regel" .

Es sei N die Gesamtzahl der Messungen des Datensatzes. Dann gilt:

$$a_{Wahr} = \bar{a} + \beta\sigma^S_{MW} \qquad \boxed{\sigma^S_{MW} = \frac{\sigma_{N-1}(a)}{\sqrt{N}}} \qquad !!!!!!!$$

wobei β die Größenordnung 1 hat.

(6.3.5) Das so entstehende Endresultat einer Messreihe wird in der Regel in folgender Form verkürzt angegeben bzw. dargestellt:

$$a = \underbrace{3.71}_{\text{für } \bar{a}} \pm \underbrace{0.05}_{\text{Für } \Delta a}$$

Man gibt also einen Bereich an, in dem - bzw. in dessen Nähe - der wahre Wert liegen sollte.

Einige Erläuterungen:

$\Diamond$ (6.3.6) Die Messungen liefern zunächst die Datenstreuung σ_{N-1} und den Schätzwert $\bar{a}$ für den Mittelwert. Die Genauigkeit der Vorhersage hängt aber nicht nur von der Datenstreuung σ_{N-1} ab, sondern auch von der Anzahl N der Messungen. Der gesamte Datensatz enthält mehr Information als die beiden Zahlen und der wahre Wert liegt daher näher am Mittelwert als durch σ_{N-1} beschrieben. Das oben bestimmte σ^S_{MW} **ist der aus dem Datensatz folgende Schätzwert für den Fehler**, die Abweichung der Mittelwerte vom wahren Mittel.

Man bestimmt also zunächst $\bar{a}$ und σ_{N-1} (N-1-Regel!) und damit dann $\sigma^S_{MW} = \frac{\sigma_{N-1}(a)}{\sqrt{N}}$ nach der $\sqrt{N}-$Regel und fasst das Ergebnis in der oben gegebenen Form zusammen.

$\Diamond$ (6.3.7) Division durch N-1 statt N **vergrößert** die Varianz leicht. Je kleiner die Datenzahl N ist, umso größer ist der Unterschied beider Werte. Die üblichen Taschenrechner bestimmen beide Streuungen (um das Datenmittel und um den wahren Wert), meist mit Bezeichnungen σ_n und σ_{n-1}. Nur die eventuell noch erforderliche weitere Division durch $\sqrt{N}$ muss man dann noch zusätzlich ausführen.

$\Diamond$ (6.3.8) In dem in (6.2.8) angegebenen Datensatz war die Datenstreuung $\sigma_n = 83$. Dagegen ergibt die Schätzung der Streuung um den wahren Wert $\sigma_{n-1} = 87$. Das ergibt als Schätzung für die Abweichung des wahren Wertes vom Mittelwert $\bar{a} = 806$ gerade $\sigma_{Mw} = \frac{87}{\sqrt{10}} = 28$. Das ist deutlich kleiner.

(6.3.9) Wieso haben wir σ^S_{MW} auch "Streuung der Mittelwerte" genannt? Mehrere Mittelwerte? Wo sollen die herkommen? Unser Datensatz liefert nur einen. Dazu stellen wir uns vor, dass wir die Messreihe, die uns unseren Datensatz $(a_1, ..a_N)$ lieferte mehrfach wiederholen. Das gibt neue Datensätze mit neuen Daten und auch jeweils mit einem zugehörigen Mittelwert. Zusammen gibt das einen Datensatz von Mittelwerten. Von diesem bilden wir den "Mittelwert der Mittelwerte" und die "Streuung der Mittewerte". Letztere ist unser σ_{MW}. wie die mathematische Problemanalyse zeigt. Nochmals die Formel, wobei die Gleichheit im Sinne von $\approx$ zu verstehen ist:

$$\sigma_{MW} \approx \sigma^S_{MW} = \frac{\sigma_{Daten}}{\sqrt{N}}$$

Das erlaubt folgende bemerkenswerte Interpretation: Die rechte Seite kann mit einer einzigen Messreihe bestimmt werden. Die linke dagegen erst nach mehrfacher Wiederholung. D.h. wir können mit einer Reihe ein Ergebnis mehrfacher Reihenwiederholung (angenähert) **vorhersagen**.

$\square$ (6.3.10) Die Formel sagt auch voraus, wie sich die Vorhersagegenauigkeit mit Vergrößerung von N, der Anzahl der Messungen, verbessert. Dabei bleibt σ_{Daten} weitgehend konstant. Angenommen Sie wollen die Genauigkeit einer Messung um eine Kommastelle verbessern, indem Sie N vergrößern, also mehr Messungen machen. Um wieviel müssen Sie dazu N vergrößern? Konkret: Sie starten mit N=10. Wie häufig müssen Sie messen, wenn Sie (bei nur statistischem Fehler) eine Stelle mehr vorhersagen wollen?

6.4 **Beispiele für die Regeln**

Es folgt jetzt je ein illustrierendes Beispiel für die beiden wichtigen Regeln: Die 1/(N-1)-Regel für die Schätzung der Abweichung vom wahren Wert und die $1/\sqrt{N}$-Regel zur Bestimmung der Streuung der Mittelwerte.

Die Beipiele werden und sollen verdeutlichen, dass beide Regeln das Behauptete leisten und wie sie das tun.

■(6.4.1) 1. Beispiel: Schätzung der **Abweichung der Daten vom wahren Wert .**

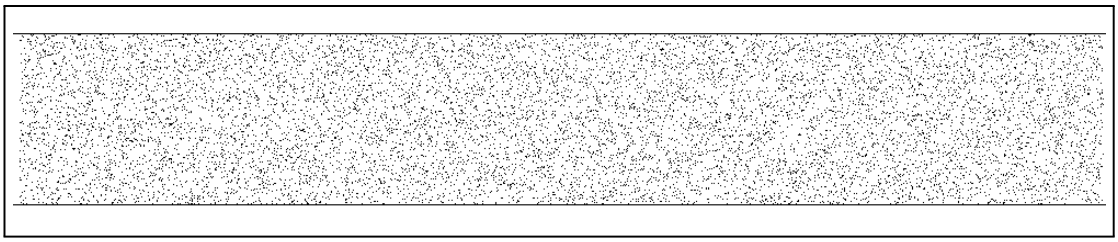
Die 1/N-Formel beschreibt die Abweichung vom Datenmittel, die 1/(N-1)-Formel schätzt die Abweichung vom wahren Wert..

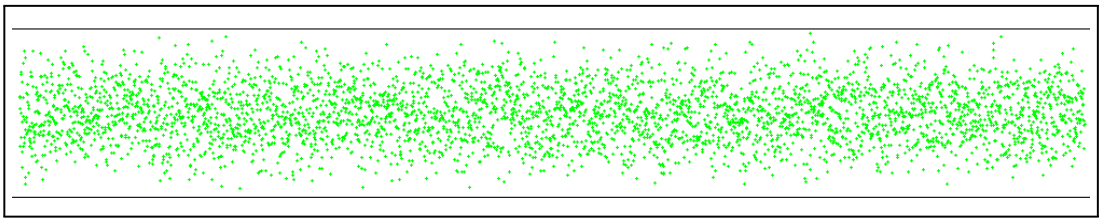
Dieser Sachverhalt soll durch das erste Beispiel verdeutlicht werden.

(6.4.2) Wir verdeutlichen den zugehörigen Sachverhalt mit einer **Computersimulation** wie folgt:

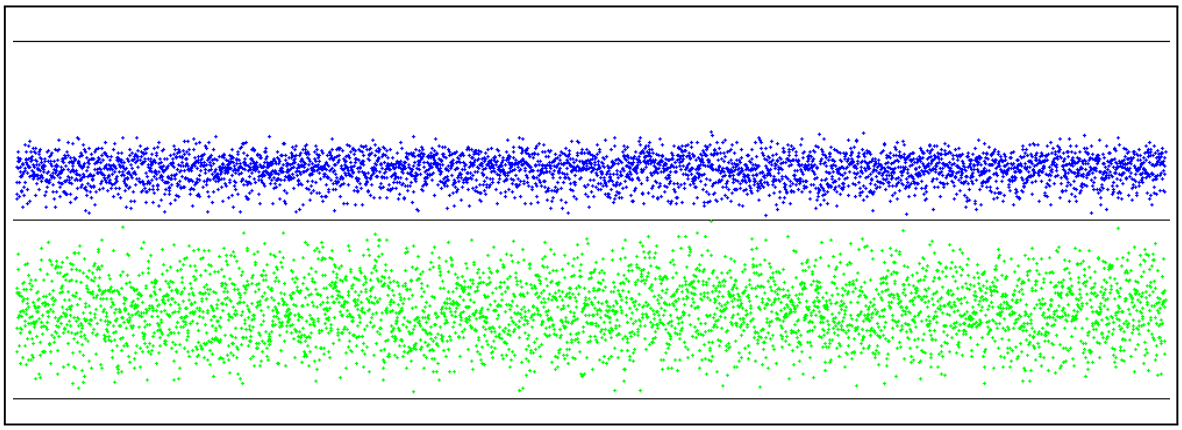
Wir wählen drei Zahlen zufällig und gleichverteilt zwischen 0 und 1 aus und bilden das arithmetische Mittel dieser drei Zahlen. (Etwa $a_1 = 0.25$, $a_2 = 0.83$ und $a_3 = 0.61$ mit $\bar{a}$=0.56.)

Das tun wir 4000 Mal. Alle diese Mittel liegen zwischen 0 und 1. Die erste Figur zeigt die Daten selbst, horiontal die Nummer der Ziehung, vertikal die Größe der zwischen 0 und 1 liegenden Zahlwerte. Im zweiten Bild sind die Mittelwerte der 4000 Dreierdatensätze grün aufgetragen. Diese Mittelwerte liegen zwischen 0 und 1, aber sie häufen sich in der Mitte. Der Grund sollte klar sein: Die Mittelwerte liegen näher am wahren Wert als die Einzeldaten.



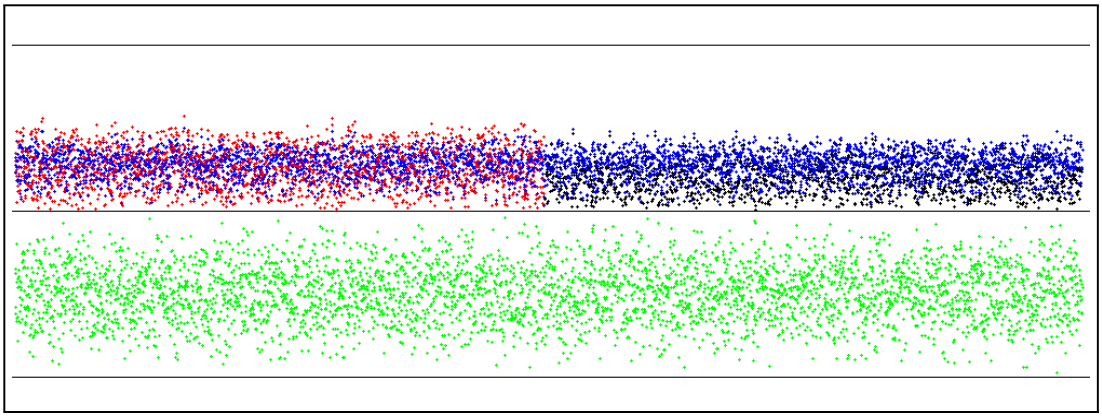


(6.4.3) Der "wahre Mittelwert" ist hier (bei Gleichverteilung) natürlich 0.5. Denn wenn man viele (einige hundert) solcher Zahlen zufällig zieht, wird deren Mittel gegen diesen Wert streben. Wir bestimmen jetzt für jeden der 4000 Datensätze die Streuung um den wahren Wert, also die Größe σ_{Wahr} aus (6.3.1). Das beschreibt den .typischen **Unterschied zwischen den Daten und dem wahren Mittel 0.5.** Das nächste Bild zeigt, wie das in unserem Computerexperiment aussieht:



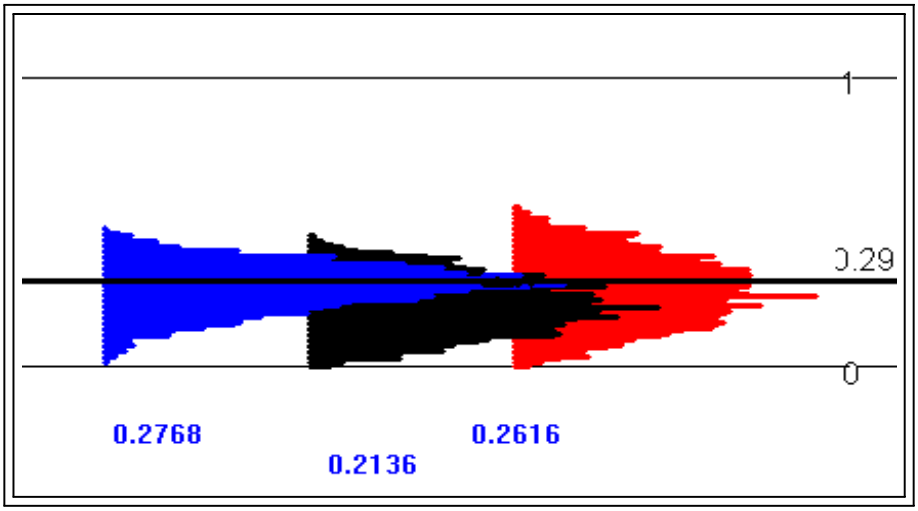
Oben ist blau dieses σ_{Wahr} für die 4000 Datensätze aufgetragen mit Wert zwischen 0 (unterer Strich) und 1 (oberer Strich). Man sieht, dass sich so relativ kleine Werte (gegenüber 1) ergeben. Aber beachten Sie: Die blauen Punkte können wir nur berechnen, weil wir im Beispiel den wahren Wert 0.5 kennen! Ein ganz kleines σ_{Wahr} bedeutet, dass alle drei gezogenen Zahlen zufällig ganz nahe bei 0.5 liegen. Wie man sieht, ist das erwartungsgemäß sehr selten!

(6.4.4) Kann man nun die Größe dieser Abweichung, also σ_{Wahr}, **mit Hilfe eines Dreierdatensatzes allein schätzen**? Das behauptete die 1/(N-1) Regel. Darum berechnen wir als nächstes einmal die Streuung der Daten mit 1/N und einmal mit 1/(N-1).



Links ist rot für die ersten 2000 Fälle die 1/(N-1)-Schätzung für σ_{Wahr} aufgetragen und man sieht, dass blau und rot etwa **dieselbe typische Größe** haben. Rechts dagegen ist schwarz die mit 1/N berechnete Datensatzstreuung σ_{Daten} aufgetragen. Und dieise Werte sind offensichtlich systematisch zu klein! **Die schwarzen Punkte liegen zu tief, würden im Mittel eine zu kleine Schätzung ergeben**.

(6.4.5) Jetzt nehmen wir noch Auszählungen all dieser Größen vor (Datensatz ist die beschriebene 4000-fache Wahl von drei Zufallszahlen). Die Auszählungen werden als Histogramme dargestellt, wobei allerdings die Bin-Werte vertikal, die Anzahlen horizontal aufgetragen sind. Die vertikale Linie liegt bei $\sigma = 0.29$, dem theoretischen Wert der Streuung der Gleichverteilung.



Von links nach rechts: Blau die Abweichungen vom wahren Mittel 0.5, also die Verteilung der σ_{Wahr}. Der Mittelwert ist 0.28. Daneben schwarz die nach der 1/N - Regel geschätzte Streuung mit der mittleren Vorhersage (über die 4000 Fälle) von 0.22. Dieser Wert ist zu klein. Rot daneben die mit der 1/(N-1)-Regel gefundene Schätzung mit dem viele besseren Wert von 0.27.

(6.4.6) **Fassen wir zusammen:** Kennt man nur eine einzige Messreihe von drei Werten (a_1, a_2, a_3), dann ergibt deren arithmetisches Mittel $\bar{a}$ den besten Schätzwert für den wahren Wert. Wie weit aber der (dann unbekannte) die Einzeldaten und der wahre Wert voneinander entfernt sind, sollte man mit der 1/(N-1)-Regel schätzen, in unserem Fall also über die Formel

$$\sigma^S_{Wahr} = \sigma_2(a) = \sqrt{Var_2(a)} \qquad \mathrm{Var}_2(a) = \frac{1}{2}\left((a_1-\bar{a})^2 + (a_2-\bar{a})^2 + (a_3-\bar{a})^2\right)$$

Ist N größer, sagen wir 10, dann ergeben die beiden Formeln für σ_N und σ_{N-1} kaum einen Unterschied. Der Unterschied wird nur für ganz kleine N beachtenswert.

Mit Hilfe der folgenden Näherungsformel

$$\frac{1}{\sqrt{N-1}} - \frac{1}{\sqrt{N}} = \frac{1}{\sqrt{N}}\left(\left(1-\frac{1}{N}\right)^{-\frac{1}{2}} - 1\right) \approx \frac{1}{\sqrt{N}} \cdot \frac{1}{2}\frac{1}{N}$$

erhält man eine Vorstellung von der Größe der N-Abhängigkeit. Denn es gilt ja:

$$\begin{aligned}
\sigma_{N-1} - \sigma_N &= \left(\frac{1}{\sqrt{N-1}} - \frac{1}{\sqrt{N}}\right)\sqrt{\sum_{i=1}^{N}(a_i-\bar{a})^2} \\
&= \sqrt{N}\left(\frac{1}{\sqrt{N-1}} - \frac{1}{\sqrt{N}}\right)\cdot\sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i-\bar{a})^2} \\
&= \sqrt{N}\left(\frac{1}{\sqrt{N-1}} - \frac{1}{\sqrt{N}}\right)\cdot\sigma_N \approx \frac{1}{2N}\sigma_N
\end{aligned}$$

Das gibt für den relativen Unterschied der beiden Größen:

$$\boxed{\frac{\sigma_{N-1}-\sigma_N}{\sigma_N} \approx \frac{1}{2N}}$$

Für N=10 also etwa 5% Unterschied.

■(6.4.7) Eine Erweiterung des Simulationsmodelles zur Illustration **der Abweichung des Mittelwertes vom wahren Wert.**

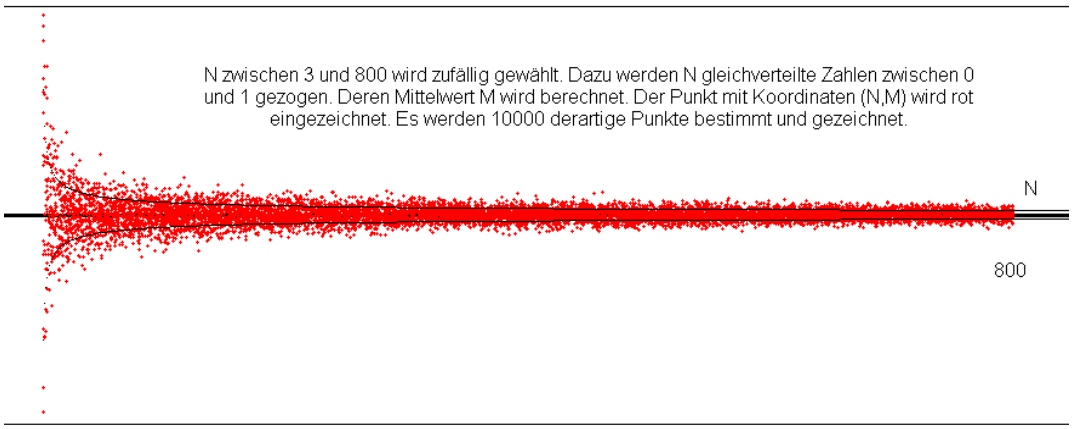
Unser Beispiel hat bisher die "(N-1)-Regel" zur Bestimmung der **Streuung der Datenwerte um den wahren Wert** erläutert. Man kann es nun aber so modifizieren, dass es auch die zweite Regel zur Bestimmung der "**Streuung der Mittelwerte**" illustriert. Diese besagte:

$$\sigma_{MW} = \frac{\sigma_{Daten}}{\sqrt{N}}$$

Besitzt man mehrere Datensätze derselben Art, dann streuen deren Mittelwerte mit σ_{MW} um den Mittelwert der Mittelwerte. Dabei kann die rechte Seite bereits mit Hilfe eines einzigen Datensatzes geschätzt werden.

(6.4.8) Jetzt die Erweiterung unseres Beispieles: **Wir ziehen nicht 3 zwischen 0 und 1 liegende zufällige Zahlen wie bisher, sondern N Stück und bilden deren Mittelwert**. Dann sollte das Resultat gemäß der gegebenen Formel um den wahren Mittelwert $\frac{1}{2}$ streuen. Wiederholt man das für ein und dasselbe N vielfach, dann sollte grob etwa die Hälfte dieser Mittelwert im Bereich $\boxed{\frac{1}{2} - \sigma_{MW} \leq \bar{a} \leq \frac{1}{2} + \sigma_{Mw}}$ liegen, die anderen außerhalb. Mit wachsendem N wird dieser Bereich wegen des Faktors $1/\sqrt{N}$ immer kleiner und zwar in einer ganz bestimmten Weise: Da für die gegebene Gleichverteilung σ_{Daten} konstant ist und zwar etwa 0.3 sollte diese Datenmitte etwa bei $\frac{0.3}{\sqrt{N}}$ liegen. Vgl. (6.2.5).

(6.4.9) Das können wir leicht mit Hilfe des Computers testen. Das Bild zeigt das Resultat eines derartigen Computerexperimentes. Beachten sie: Horizontal ist jetzt N aufgetragen, nicht etwa die Nummer der Ziehung.



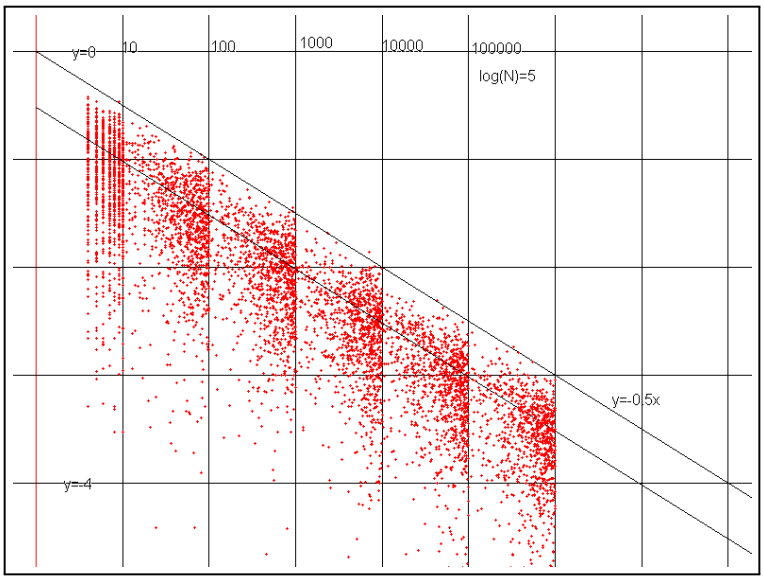
Zum Vergleich sind die beiden Kurven M=0.5$\pm\frac{0.3}{\sqrt{N}}$ mir eingezeichnet! Man sieht, dass die behauptete N-Abhängigkeit grob erfüllt ist.Und man sieht, dass die Streuung mit N nur sehr langsam kleiner wird! Nur: Wenn man N noch weiter vergrößert, wird das Bild immer weniger aussagekräftig.

(6.4.10) Um ein deutlicheres Bild zu erhalten, ist eine **logarithmische Auftragung der Datenpunkte** vorzuziehen. Aus der Gleichung $\boxed{\sigma_{MW}\text{(N)-0.5=}\frac{0.3}{\sqrt{N}}}$ folgt durch Logarithmieren

$$\boxed{\log(|\sigma_{MW}\text{-0.5}|)=\log(0.3)\text{-}\frac{1}{2}\log(\text{N})}$$

Wir lassen wie im ersten Bild den Computer ein Punktepaar (N,M) berechnen, tragen aber den Punkt $\boxed{(\log(\text{N}),\log(|\text{M-0.5}|))}$ auf und vergleichen mit der Geraden $\boxed{\text{y=log(0.3)-}\frac{1}{2}x}$ sowie mit der Geraden y=0-$\frac{1}{2}x$.

Um einen einigermaßen verteilten Punktschwarm zu erhalten wählen wir statt $N_{max} = 800$ aus dem ersten Bild zuerst $N_{max} = 10 = 10^1$, dann $N_{max} = 100 = 10^2$ usw. bis $N_{max} = 1000000 = 10^6$. Zu jedem dieser Obergrenzen bestimmen wir 1000 Punkte in der beschriebenen Weise:



Mittelwerte M, die nahe beim wahren Wert 0.5 liegen haben einen nach $-\infty$ gehenden Wert von log(|M-.5|). Genauer: Ist die Abweichung |M-0.5| größer als 0.3, dann liegt der Punkt **oberhalb** der Geraden y=log(0.3)-$\frac{1}{2}x$. Ist sie kleiner, darunter. Man sieht beispielsweise: Ab $N=10^6$ ist die Abweichung (des gefundenen vom wahren Mittel) fast immer kleiner als 10^{-3}.

Auszählen zeigt: Etwa 30% der Punkte liegt über der Referenzgeraden y=log(0.3)-0.5log(x). Kaum einer liegt über y=-.5log(x). Das zeigt den Gehalt des $\frac{1}{\sqrt{N}}$-Gesetzes für unser Beispiel.

■ (6.4.11) Das 2. Beispiel: zur Illustration der **Abweichung des Mittelwertes vom wahren Wert.** Es soll die praktische Nützlichkeit der zweiten Regel - der $1/\sqrt{N}-$ Regel zur Schätzung des wahren Wertes - verdeutlichen.

(6.4.12) Wir beginnen mit der Erstellung eines Datensatzes vom *Auszählungstyp:*

♦ Wähle eine Buchseite. Numeriere die Zeilen durch. Zähle wieviel Worte in der i-ten Zeile stehen. Bezeichnung W_i. Dann ist $(W_1, ..., W_N)$ ein typischer Datensatz vom Auszählungstyp. N ist die Zahl der Zeilen auf der Seite. Jetzt wird diese Liste nach der Anzahl ausgezählt. n_k bezeichnet die Anzahl der Zeilen auf der Seite mit k Worten.

(6.4.13) **Jetzt überlegen wir wie folgt:** Eigentlich interessiert uns die mittlere Wortzahl pro Zeile für das **gesamte Buch**. Dazu müssten wir alle Seiten auszählen. Dann gäbe der entstehende Datensatz die gesuchte Größe ebenso wie die zugehörige Streuung.

Aber was sagen die Daten der einen ausgezählten Seite bereits über diese Größe aus? Vermutlich liegt das Gesamtmittel in der Nähe des Seitenmittels. Aber wie groß könnte der Fehler sein, die Abweichung der beiden Größen voneinander? Die erste Idee wäre, die (über die Auszählung einer Seite) bestimmte Datenstreuung als Maß zu nehmen. Aber dieser Wert ist zu groß! Nach (4.3.4) sollten wir vielmehr diese Größe noch durch $\sqrt{N}$ teilen, wenn N die Zahl der Zeilen der Seite ist.

Experimentell können wir so vorgehen: Wir zählen nicht eine Seite aus, sondern vielleicht 5 oder 10. Für jede Seite bestimmen wir Mittelwert und Streuung. Das gibt uns einen neuen Datensatz mit 10 Mittelwerten $(\bar{a}_1, \bar{a}_2, ..., \bar{a}_{10})$. Weiter finden wir 10 Seitenstreuungen, die alle ungefähr gleich sind. **Für diesen neuen Datensatz der Mittelwerte** bestimmen wir wieder Mittel und Streuung. Und diese neue Streuung, die Streuung der Mittelwerte, sollte etwa um den Faktor $1/\sqrt{N}$ kleiner als die Seitenstreuung sein. Diese Vorhersage können wir testen!.

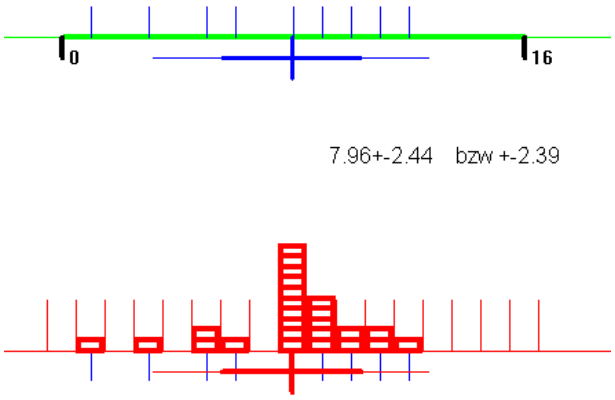


Figure 1:

Also: Wir zählen eine Seite aus und berechnen Seitenmittelwert $\bar{a}$ und Streuung σ_{Seite}. Dann bilden wir $\frac{\sigma_{Seite}}{\sqrt{N}}$ und das ist eine Schätzung für den Fehler, den man macht, wenn man statt des Buchmittels das Seitenmittel nimmt. Oder formal

$$\boxed{\bar{a}_{Buch}=\bar{a}_{Seite}+\frac{\sigma_{Seite}}{\sqrt{N}}\beta} \quad \text{mit } |\beta| \text{ in der Größenordnung 1}$$

Die Zahl β ist natürlich nicht bekannt, aber ihre Größenordnung. Für N=30 (Zeilen pro Seite) ist daher die Vorhersage für den wahren Mittelwert etwa um den Faktor $\frac{1}{\sqrt{30}} = 0.182$ genauer als die Seitenstreuung

(6.4.14) Ein zugehöriges konkretes Zahlbeispiel: Die Tabelle enthält die Wörterzahl der ersten 25 Zeilen einer Buchseite. Damit ist $\sqrt{25} = 5$.

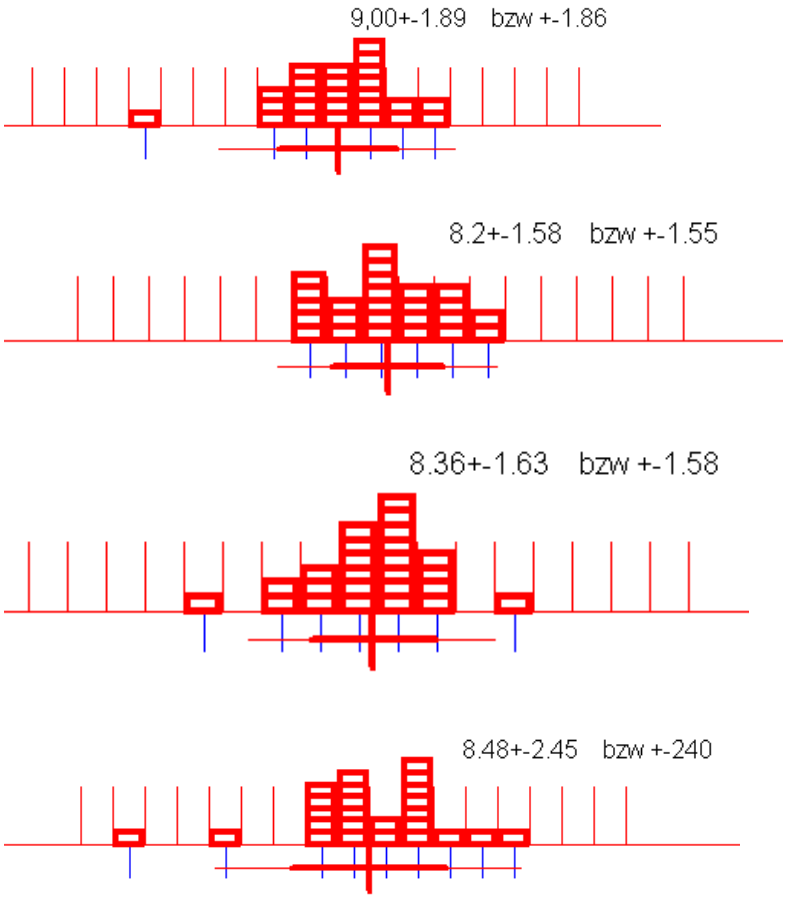
9	8	9	10	8	9	11	9	8	8	5	1	10	8	8	11	5	8	8	12	8	6	9	3	8

Die Figur zeigt zwei graphische Darstellungen dieses Datensatzes, als Werteverteilung auf der Zahlgeraden und als Auszählungshistogramm.

Datenmittelwert und Datenstreuung sind mit eingezeichnet. Man erhält als Datensatzbeschreibung nach 6.2 $\boxed{7.96\pm2.39}$.

(6.4.15) Nach unseren Regeln gibt das eine Vorhersage von $\boxed{7.96\pm0.49}$ für den "wahren Wert". Das ist in diesem Fall die mittlere Wörterzahl pro Zeile für das gesamte Buch. Oder etwas einfacher für - sagen wir - für 5 Seiten.

(6.4.16) Jetzt zählen wir weitere 4 Seiten aus. Nachfolgend die Histogramme. Man sieht, dass beträchtliche Formunterschiede vorliegen mit einer Streuung jeweils von der Größenordnung 2.



(6.4.17) Die Auswertung jeder Einzelseite macht eine Vorhersage $\bar{a}_i \pm \frac{\sigma_{Di}}{5}$ für den wahren Wert. Dabei bezeichnet $\bar{a}_i$ den Mittelwert und σ_{Di} die Datenstreuung der i-ten Seite. Wir finden

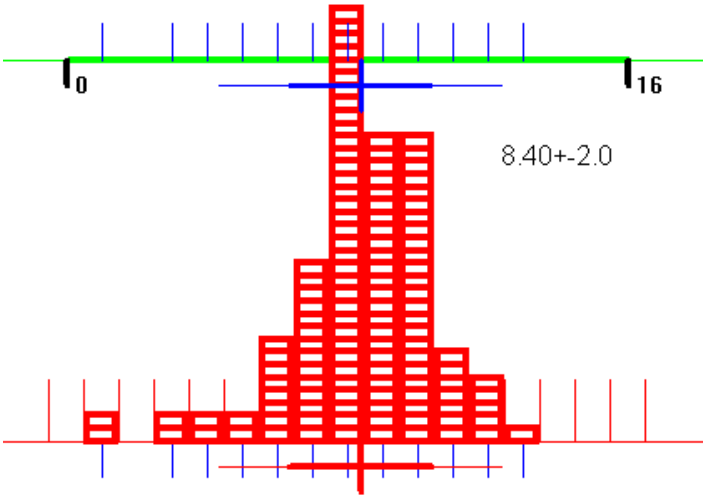
7.96 ± 0.49	9 ± 0.38	8.2 ± 0.32	8.36 ± 0.33	8.48 ± 0.49

(6.4.16) Der Datensatz der Mittelwerte liefert das folgende Bild (mit Gesamtmittel und Mittelwertstreuung).

Wir sehen: Bei 4 unserer Vorhersagen liegt der "wahre Wert (über die 5 Seiten)" von 8.4 im angegebenen 1-σ − *Intervall*. In einem Fall liegt er außerhalb, aber immer noch gut im 2σ − *Bereich*.

Und: Die die Vorhersagegenauigkeit bestimmende Streuung der Mittelwerte von 0.4 ist hier deutlich kleiner als die Datenstreuung ($0.4=\frac{2}{4}$ gegen 2.0).

(6.4.17) Jetzt noch der Gesamtdatensatz der 5 Seiten. Die Verteilung ist bereits viel regelmäßiger als die für die Einzelseiten.



☐ Bestimmen Sie die "Koordinatenwerte" β_i , die das "wahre" Gesamtmittel durch die jeweiligen Schätzgrößen ausdrücken, für die also gilt

$$8.4 = \bar{a} = \bar{a}_i + \frac{\sigma_{Di}}{5}\beta_i$$

Das sind ja die beim üblichen Messvorgang unbekannt bleibenden Werte, von denen man nur die Größenordnung kennt!

6.5 Lineare Regression

★ Wie sollte man vorgehen, um einen **Text wie den ab (6.5.1) gegebenen zu erarbeiten**?

♦♦ Zunächst das anstehende Problem verstehen. Das sollte hier durch genaue Lektüre der ersten beiden Absätze (6.5.1) erfolgen! Eventuell noch das Beispiel (6.5.4) und die dortige Figur als Erläuterung anschauen. Jetzt sollten Sie sich folgende Fage stellen und beantworten:

♦ Was ist gegeben, was gesucht und was müssen Formeln leisten, die die Antwort liefern?

◊ Fallen Ihnen Beispiele ein, in denen das gestellte Problem auftreten könnte?

♦♦ Jetzt inspizieren sie den Text, ob Sie dort Formeln der gesuchten Art finden. Die Antwort finden Sie sicher bereits über die graphische Textgestaltung. Es sind die Formeln aus (6.5.3).

♦ Leisten diese Formeln das Gewünschte? Verstehen Sie die selbsterklärenden Bezeichnungsweisen? Notfalls das Beispiel zur Hilfe nehmen.

◊ Zur Übung eventuell ein eigenes Beispiel rechnen.

♦♦ Jetzt können Sie sich fragen: Benötigen Sie die Herleitung der Formeln? Möchten Sie diese verstehen, den Beweis nachvollziehen?

◊ Wenn **ja**, den "Text dazwischen" durchgehen und bearbeiten. Also (6.5.2).

◊ Danach sollten Sie die Frage nach der "Beweisidee" beantworten können. (Die Formel für A(m,b) erläutern, möglichst graphisch. Wie sind m und b zu wählen?)

◊ Wenn **nein**, lieber noch etwas über die Unzulänglichkeiten der Problemlösung nachdenken, von denen es einige gibt. Das können Sie natürlich auch nach Durchgehen des Beweises tun. Insbesondere zeigt die Beweisidee gewisse Unzulänglichkeiten des Resultates auf.

☐ Können Sie m und b mit Hilfe Ihres Taschenrechners bestimmen?

☐ Inwiefern kann man die hier behandelten Typ von Datensätzen auch als Zeitreihendsätze interpretieren?

Der Text

■ (6.5.1) Wir haben einen Punkteschwarm von N Punkten in der Koordinatenebene. Also

$$(x_1, y_1), (x_2, y_2)...., (x_N, y_N).$$

Dabei sollen die $x_i - Werte$ exakt sein, die y_i-Werte dagegen dürfen Fehler und Ungenauigkeiten enthalten.

Eigentlich (1,(x_1, y_1)).(2, (x_2, y_2)), ..., (N, (x_N, y_N)). Wir benutzen wieder die verkürzte Darstellung, die nur die Werte angibt.

Wir wollen eine Gerade mit Gleichung $\boxed{y=mx+b}$ durch diesen Schwarm legen, die diesen Schwarm besonders genau wiedergibt. Wir setzen $Y_i = mx_i + b$. Dann soll die Gesamtheit der Y_i das jeweilige y_i besonders gut approximierern. Genauer. Die Abweichungen $y_i - Y_i$ zwischen dem Datenwert y_i und dem zugehörigen Punkt auf der Geraden soll **insgesamt** möglichst klein werden. Damit sich Beiträge mit unterschiedlichem Vorzeichen in der Summe nicht fortheben, bilden wir $(y_i - Y_i)^2$ und summieren. Das gibt die Größe

$$A(m,b) = \sum_{i=1}^{N} (y_i - Y_i)^2 = \sum_{i=1}^{N} (y_i - mx_i - b)^2$$

Unser Problem: F**ür welche Wahl von m und b wird A minimal?? Das sollte die bestapproximierend Gerade sein!** Beachten Sie: Ist der Datensatz gegeben und ebenso m und b, dann kann man den Wert A(m,b) bestimmen.

(6.5.2) Man überlegt sich leicht, dass es mindestens ein solches Minimum geben muss. Für dieses müssen die beiden partiellen Ableitungen verschwinden. Oder auch: Die momentanen Änderungsraten in m und b müssen Null werden. Ableiten gibt:

$$\begin{aligned} \frac{\partial A}{\partial m}(m,b) &= -2\sum_{i=1}^{N} x_i(y_i - mx_i - b) \stackrel{!}{=} 0 \\ \frac{\partial A}{\partial b}(m,b) &= -2\sum_{i=1}^{N} (y_i - mx_i - b) \stackrel{!}{=} 0 \end{aligned}$$

Das sind zwei Bedingungen zur Bestimmung der beiden Unbekannten m und b. Auswerten der zweiten Summe gibt:

$$\sum_{i=1}^{N} (y_i - mx_i - b) = \sum_{i=1}^{N} y_i - m\sum_{i=1}^{N} x_i - \underbrace{N}_{\sum 1} b = 0$$

Andererseits ist $\boxed{\sum_{i=1}^{N} y_i = N\bar{y}}$ und $\sum_{i=1}^{N} x_i = N\bar{x}$. Der Vektor ($\bar{x}$,$\bar{y}$) ist der Schwerpunktsvektor des gesamten Schwarmes! Oder $\boxed{\bar{y}\text{-}m\bar{x}\text{-}b=0}$. D.h. die Lösungsgerade geht durch den Schwerpunkt des Schwarmes!

Jetzt formen wir die erste Bedingungsgleichung entsprechend um und finden

$$\sum_{i=1}^{N} x_i y_i - m\sum_{i=1}^{N} x_i^2 - b\sum_{i=1}^{N} x_i \stackrel{!}{=} 0$$

$$\boxed{(\vec{x}\cdot\vec{y}) = m\vec{x}^2 + b\bar{x}} \quad \text{mit} \quad \boxed{\sum_{i=1}^{N} x_i y_i = N(\vec{x}\cdot\vec{y})} \quad \text{und} \quad \boxed{N\vec{x}^2 = \sum_{i=1}^{N} x_i^2}$$

Damit haben wir zwei einfache Gleichungen für die beiden Unbestimmten m und b:.

$$\begin{aligned} (\vec{x}\cdot\vec{y}) &= m\vec{x}^2 + b\bar{x} \\ \bar{y} &= m\bar{x} + b \end{aligned}$$

(6.5.3) Auflösen gibt das folgende Resultat:

$$\boxed{m = m_{reg} = \frac{(\vec{x}\cdot\vec{y}) - \bar{x}\bar{y}}{\vec{x}^2 - \bar{x}^2}} \quad \text{mit} \quad N\bar{x} = \sum x_i \quad N\bar{y} = \sum y_i$$

$$\boxed{b = b_{reg} = \frac{\vec{x}^2\bar{y} - (\vec{x}\vec{y})\bar{x}}{\vec{x}^2 - \bar{x}^2}} \quad \text{und mit} \quad N(\vec{x}\cdot\vec{y}) = \sum x_i y_i \text{ und } N\vec{x}^2 = \sum x_i^2$$

Klar: Hat man einen Datensatz der erforderlichen Art, dann kann man für $\vec{x}^2 \neq \bar{x}^2$ die beiden Größen m und b berechnen und von Sonderfällen abgesehen ergibt das eine angemessene Gerade.

Allerdings sollte man Datensatz und Gerade wenn irgend möglich graphisch darstellen und das Resultat inspizieren, ob es die Daten sinnvoll beschreibt.

Von der Herleitung benötigt man nur die Idee.

(6.5.4) **Beispiel:** Der Datensatz sei $\boxed{(1,1.2),(2,1.8),(2,2.3),(3,3.1),(4,4.3)}$ mit N=5. Denken sie daran: (1,1.2) ist der erste Wert im Datensatz und steht hier für (1,(1,1.2)).

Das gibt den Schwerpunktsvektor:

$$\begin{aligned} 5\vec{x}_S &= 5(\bar{x},\bar{y}) = (1,1.2)+(2,1.8)+(2,2.3)+(3,3.1)+(4,4.3) \\ &\boxed{5\vec{x}_S=(12,12.7)} \end{aligned}$$

$$5(\vec{x}\cdot\vec{y}) = \begin{pmatrix}1\\2\\2\\3\\4\end{pmatrix}\cdot\begin{pmatrix}1.2\\1.8\\2.3\\3.1\\4.3\end{pmatrix} = 35.9$$

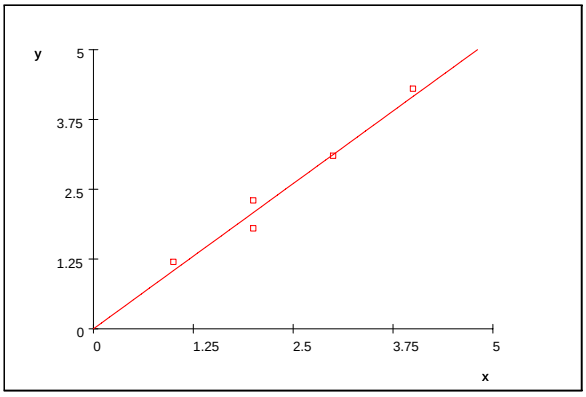
$$5\vec{x}^2 = \begin{pmatrix}1\\2\\2\\3\\4\end{pmatrix}\cdot\begin{pmatrix}1\\2\\2\\3\\4\end{pmatrix} = 34$$

Nun können wir einsetzen und finden

$$\begin{aligned} m &= \frac{(\vec{x}\cdot\vec{y})-\bar{x}\bar{y}}{\vec{x}^2-\bar{x}^2} = \frac{\frac{35.9}{5}-\frac{12}{5}\frac{12.7}{5}}{\frac{34}{5}-\frac{12^2}{5^2}} = 1.0423 \\ b &= \frac{\vec{x}^2\bar{y}-(\vec{x}\vec{y})\bar{x}}{\vec{x}^2-\bar{x}^2} = \frac{\frac{34}{5}\frac{12.7}{5}-\frac{35.9}{5}\frac{12}{5}}{\frac{34}{5}-\frac{12^2}{5^2}} = 3.8462\times10^{-2} \end{aligned}$$

Die beiden Werte entsprechen der Vorerwartung, die einem eine Inspektion des Datensatzes liefert.

Als Figur:



(6.5.5) Ein **Vertrauenstest** für die beiden hergeleiteten Formeln zur Bestimmung der Regressionsgeraden. Sagen wir N=10^6 (Distraktor) wobei $\boxed{y_i = mx_i + b}$ für alle i sein soll.

Vorerwartung? m_{reg} =?? b_{reg} = ..? (Läßt sich klar beantworten!)

Zur Verdeutlichung bezeichnen wir die sich aus dem Datensatz ergebenden Größen neu:

$\boxed{m_{reg} = \frac{(\vec{x}\cdot\vec{y})-\bar{x}\bar{y}}{\vec{x}^2-\bar{x}^2}}$ mit $\mathrm{N\bar{x}}=\sum x_i$ $\mathrm{N\bar{y}}=\sum y_i$

$\boxed{b_{reg} = \frac{\vec{x}^2\bar{y}-(\vec{x}\vec{y})\bar{x}}{\vec{x}^2-\bar{x}^2}}$ und mit $\mathrm{N}(\vec{x}\cdot\vec{y}) = \sum x_i y_i$ und $\mathrm{N}\vec{x}^2 = \sum x_i^2$

Jetzt berechnen wir die benötigten Hilfsgrößen:

- $\boxed{\mathrm{N\bar{x}}=\Sigma x_i}$ $\mathrm{N\bar{y}}=\Sigma(mx_i+b) = mN\bar{x}+bN$. Also : $\boxed{\bar{\mathrm{y}}=\mathrm{m\bar{x}+b}}$
- $\mathrm{N}(\vec{x}\cdot\vec{y}) = \Sigma x_i y_i = \Sigma x_i(mx_i+b) = m\Sigma x_i^2 + bN\bar{x} = mN\vec{x}^2 + bN\bar{x}$. Also

$\boxed{(\vec{x}\cdot\vec{y}) = m\vec{x}^2 + b\bar{x}}$

Einsetzen gibt das erwartete Resultat:

$$\boxed{\mathrm{m}_{reg} = \frac{\left(m\vec{x}^2+b\bar{x}\right)-\bar{x}(m\bar{x}+b)}{\vec{x}^2-\bar{x}^2} = \frac{m(\vec{x}^2-\bar{x}^2)}{\vec{x}^2-\bar{x}^2} = m}$$

$$\boxed{\mathrm{b}_{reg}=\frac{\vec{x}^2(m\bar{x}+b)-(m\vec{x}^2+b\bar{x})\bar{x}}{\vec{x}^2-\bar{x}^2}=\frac{b(\vec{x}^2-\bar{x}^2)}{\vec{x}^2-\bar{x}^2}=\mathrm{b}}$$

Die Regressionsgerade stimmt mit der Geraden überein, von der alle Datenpaare genommen wurden. Der Vertrauenstest ist bestanden!.

6.6 Die Genauigkeit einer Quotenschätzung

■ (6.1.1) Hier geht es um folgendes Problem: Eine bestimmte relative Häufigkeit, also eine Quote, soll geschätzt werden. Man verfüge ober N zufällige Werte der Größe und finde, dass davon n Fälle einschlägig sind. Dann ist $\mathrm{q}_S = \frac{n}{N}$ ein Schätzwert dieser (wahren) Quote q_W. Kann man wieder näherungsweise etwas über die Genauigkeit dieser Schätzung sagen?

(6.1.2) Eine Antwort kann als einfache Regel in Form eines 1-σ – *Intervalles* gegeben werden:

$$\begin{aligned} n_W &= n \pm \sqrt{n} \\ &\boxed{\mathrm{q}_W=\tfrac{n}{N} \pm \tfrac{\sqrt{n}}{N}} \end{aligned}$$

☐ Leistet die Formel das Gewünschte?

☐ Überlegen sie sich Anwendungsbeispiele.

☐ Denken sie sich ein Computerexperiment aus, mit dessen Hilfe man diese Regel testen und verdeutlichen kann.
