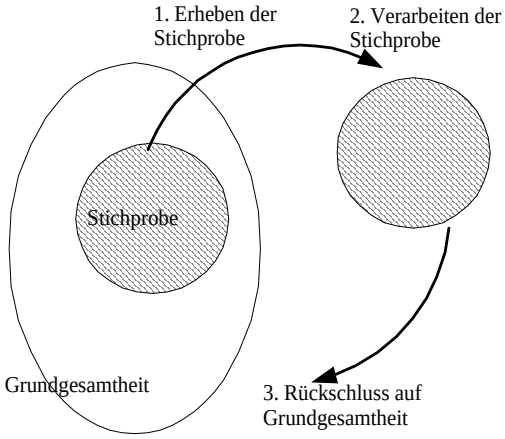
EINFÜHRUNG IN METHODEN DER EIN- UND ZWEIDIMENSIONALEN STATISTIK (MICHAEL SPIELMANN)

Einführung
in Methoden der ein- und zweidimensionalen Statistik
(Michael Spielmann)

Eindimensionale Stichproben

1. Aufgabe der eindimensionalen Statistik und grundlegende Begriffe

Grundgesamtheit, Stichprobe, Kennzahlen
In vielen Bereichen der Wissenschaft und Technik will man mit Hilfe der **deskriptiven Statistik** Aussagen über eine **Grundgesamtheit** machen, die zu groß ist, um sie ganz zu erfassen. Man untersucht stattdessen eine **Stichprobe**. Sie wird geordnet, klassiert, graphisch veranschaulicht, zahlenmäßig ausgewertet. Dazu berechnet man Häufigkeiten, Mittelwerte, Streuungen.
Die **absolute Häufigkeit** eines Ereignisses gibt an, wie oft das Ereignis tatsächlich eingetreten ist. Die **relative Häufigkeit** benötigt man als Vergleichszahl bei Serien oder Zufallsexperimenten unterschiedlichen Umfangs. Unter 100 Versuchen trat 55-mal G ein, dann ist die absolute Häufigkeit 55 und die relative Häufigkeit ist 55/100=55%=0,55. Die Tabelle, die die Stichprobenwerte und die zugeordneten relativen Häufigkeiten enthält, nennt man **Häufigkeitsverteilung**.
Der **Mittelwert** als einzelner Zahlenwert soll die Stichprobe repräsentieren. Die Information der gesamten Stichprobe wird dadurch reduziert zugunsten einer kompakten Darstellung.
Streuungen, also Abweichungen vom Mittelwert oder Spannweiten, bestimmt man, wenn man Stichproben gleichen Mittelwertes vergleichen will oder über die Struktur der Stichprobe mehr aussagen will als der Mittelwert alleine zulässt.
Mit dem Rückschluss auf die Grundgesamtheit beschäftigt sich die **Beurteilende Statistik**. Sie benötigt zur theoretischen Begründung die Wahrscheinlichkeitsrechnung.



2. Erhebung und Aufbereitung einer Stichprobe

Gewöhnlich ist eine Stichprobe in Form einer ungeordneten Liste, der sogenannten **Urliste**, gegeben. Die Anzahl der Elemente ist der **Umfang** der Stichprobe. Enthält die Urliste wenige Elemente, sortiert man sie. Enthält sie viele Elemente, fasst man sie mittels einer Strichliste in Klassen zusammen. Die Strichliste ist schon eine Häufigkeitstabelle.
Beispiel Klassenarbeit:
Urliste 2 3 1 4 3 3 4 5 2 3 2 5 4 6 4 4 3 4 2 1 3 4 3 3 2
Strichliste

1	2	3	4	5	6
II	IIIII	IIIII III	IIIII II	II	I

Man ordnet die Häufigkeitstabelle meistens vertikal an.

x_i	n_i
1	2
2	5
3	8
4	7
5	2
6	1

ANMERKUNG:
Die Stichprobenwerte haben wir x_i genannt, sie heißen auch **Merkmalswerte**.
Die Häufigkeiten haben wir n_i genannt, sie heißen auch **Besetzungszahlen**.
Man sagt auch, die Merkmalswerte seien mit den Besetzungszahlen **gewichtet**.

Zur rechnerischen Auswertung der Häufigkeitstabelle hat sich folgende Anordnung bewährt.

x_i	n_i	$x_i n_i$
1	2	2
2	5	10
3	8	24
4	7	28
5	2	10
6	1	6
Summe	**25**	**90**

3. Kennzahlen: Mittelwerte, Streuungsmaße

Definition arithmetisches Mittel

Die Information, die in der Tabelle steckt, will man komprimieren. Eine einfache Methode ist die Bestimmung eines Mittelwertes.
Arithmetisches Mittel ist der Wert, der die geordnete Stichprobe in zwei Bereiche gleichen „Gewichtes“ teilt.
Bei unklassierter Stichprobe

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_{n-1} + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Bei klassierter Probe, Einteilung in k Klassen mit Gewichten n_i

$$\bar{x} = \frac{1}{n}(x_1 n_1 + x_2 n_2 + \cdots + x_k n_k) = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$

Beispiel: Vergleich von Klassenarbeiten in einer Klasse, Fächer Deutsch-Mathematik

Beispiel

Deutscharbeit		
x_i	n_i	$x_i n_i$
1	2	2
2	5	10
3	8	24
4	7	28
5	2	10
6	1	6
Summe	**25**	**80**

$$\bar{x} = \frac{80}{25} = 3{,}2$$

Mathematikarbeit		
x_i	n_i	$x_i n_i$
1	0	0
2	4	8
3	12	36
4	9	36
5	0	0
6	0	0
Summe	**25**	**80**

$$\bar{x} = \frac{80}{25} = 3{,}2$$

Vergleich: Die Deutsch-Arbeit hat einen Mittelwert (Durchschnittsnote) von 3.2, die Mathematik-Arbeit ebenfalls. Die Arbeiten sind gleich gut ausgefallen
Ein Unterschied fällt auf. In Mathematik fehlen die Noten 1, 5 und 6, die Ergebnisse liegen enger beieinander.
Was sagt der Mittelwert aus? Wozu berechnet man ihn? Kann man die Stichprobe durch den Mittelwert zutreffend beschreiben?
Der Mittelwert ist eine durchschnittliche Note, die (wegen der Nachkommastellen offensichtlich) von niemandem erzielt wurde. Warum berechnet man ihn, wenn er nie realisiert wird? Er soll die Stichprobe repräsentieren. Er stellt einen griffigen Wert dar, der die Stichprobe beschreibt. Er ist ein Durchschnittswert, von dem die Stichprobenwerte allerdings mehr oder weniger stark abweichen.

Für einen tiefergehenden Vergleich der Stichproben benötigt man ein Abweichungsmaß, das sogenannte Streuungsmaß.

Definition Streuungsmaße, Spannweite und Varianz

Wir sahen oben, dass zwei Verteilungen sehr unterschiedliche Form haben können.
Als Streuungsmaß kann die **Spannweite** dienen. Sie ist die Differenz des größten und des kleinsten Stichprobenwertes. Die Spannweite ist allerdings nur ein grobes Maß.
Wenn in der Arbeit eine Eins und eine Sechs, sonst nur Dreien und Vieren geschrieben wurden, ist die Spannweite 5 bei möglichem Mittelwert 3,5.
Wenn in der Arbeit nur Dreien und Vieren geschrieben wurden, ist die Spannweite 1 bei möglichem Mittelwert 3,5.
Die Spannweite erfasst also nur die extremen Werte; alle anderen Werte werden nicht berücksichtigt.

Ein individuelleres Maß ist die durchschnittliche Abweichung vom Mittelwert.

Um den Sinn von Abweichungsmaßen besser zu verstehen, mache man sich die folgenden Beispiele klar.
1. Beispiel: Schießversuche
Ein Schütze kann beim Schießen auf eine Scheibe richtig schießen, das heißt, er trifft im Wesentlichen die Mitte. Dabei kann er genau oder ungenau schießen, was bedeutet, dass er eng um die Mitte oder weit verteilt trifft. Er kann aber auch falsch schießen, das heißt, er trifft nicht die Mitte, sondern einen anderen Bereich. Dabei kann er genau oder ungenau schießen, was bedeutet, dass die Treffer eng liegen oder weit verteilt. Wenn er genau schießt, muss er nur anders visieren. Wenn er ungenau schießt, muss er lange üben.
2. Beispiel: Flaschenfüllung
Mineralwasserflaschen müssen eine bestimmte Füllhöhe aufweisen. Mit der Zeit verstellt sich die Maschine. Füllt sie falsch aber genau, ist eine Korrektur notwendig. Füllt sie ungenau, wird sie gründlich überholt werden müssen.

Beispiel

gleicher Mittelwert, gleiche Spannweite, unterschiedliche « Streuung »

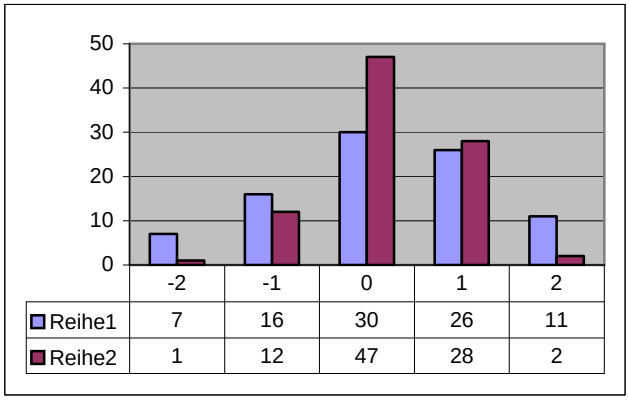
x_i	n_i	$x_i n_i$
-2	7	-14
-1	16	-16
0	30	0
1	26	26
2	11	22
Summe	**90**	**18**

$$\bar{x} = \frac{18}{90} = 0{,}2$$

x_i	n_i	$x_i n_i$
-2	1	-2
-1	12	-12
0	47	0
1	28	28
2	2	4
Summe	**90**	**18**

$$\bar{x} = \frac{18}{90} = 0{,}2$$

Hier ist ein Säulendiagramm, das sogenannte **Histogramm** informativ.



Wenn man die Summe der (linearen) Abweichungen vom Mittelwert berechnet, erhält man immer 0; positive und negative Abweichungen heben einander auf; die Vorzeichen verschwinden durch Beträge oder Quadrate. Um ein brauchbares Abweichungsmaß zu erhalten, müssen die Vorzeichen verschwinden.
Wir entscheiden uns für die Summe der quadrierten Abweichungen.

Die **mittlere Quadratische Abweichung** oder **Varianz** ist

bei unklassierter Stichprobe

$$V = s_x^2 = \frac{1}{n-1} S_{xx} = \frac{1}{n-1}\left[(x_1-\bar{x})^2 + \cdots (x_n-\bar{x})^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

bei klassierter Probe mit Einteilung in k Klassen und gewichteten Merkmalswerten

$$V = s_x^2 = \frac{1}{n-1} S_{xx} = \frac{1}{n-1}\left[(x_1-\bar{x})^2 n_1 + \cdots (x_k-\bar{x})^2 n_k\right] = \frac{1}{n-1}\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i$$

Die abkürzende Schreibweise $S_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2$ kann zu größerer Übersicht führen. Im Folgenden werden wir sie öfter benutzen.

Hinweis: Wir haben durch den Stichprobenumfang $n-1$ geteilt, um die mittlere Abweichung zu erhalten. Ein anderes übliches Verfahren ist das Dividieren durch n. Manche Taschenrechner bieten beide Streuungswerte an. Man kann nachweisen, dass der Term mit $n-1$ ein besserer Schätzwert für die Varianz ist.

Da die Varianz ein quadrierter Term ist, hat sich die Dimension verändert. Das ist nicht günstig. Wenn wir in Meter messen, nützt ein Abweichungsmaß in Quadratmeter nicht viel. Die Wurzel aus der Varianz ist als Abweichungsmaß geeigneter. Man nennt sie **Standardabweichung**

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}$$

Berechnung der Standardabweichung

x_i	n_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
-2	7	4,84	33,88
-1	16	1,44	23,04
0	30	0,04	1,2
1	26	0,64	16,64
2	11	3,24	35,64
Summe	**90**		**110,4**

$$s_x = \sqrt{\frac{110{,}4}{89}} = 1{,}114$$

x_i	n_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
-2	1	4,84	4,84
-1	12	1,44	17,28
0	47	0,04	1,88
1	28	0,64	17,92
2	2	3,24	6,48
Summe	**90**		**48,4**

$$s_x = \sqrt{\frac{48{,}4}{89}} = 0{,}737$$

Die Beispiele zeigen die Bedeutung der Standardabweichung:
ist die Standardabweichung **groß**, so sind viele Stichprobenwerte **weit** vom Mittelwert entfernt,
ist die Standardabweichung **klein**, so sind viele Stichprobenwerte **nah** um den Mittelwert geschart.
Da „groß“ oder „klein“ aber auch vom Maßstab abhängen kann, hat man den **Variationskoeffizienten** $\frac{s_x}{\bar{x}}$ definiert. So können Standardabweichungen verschiedener Variablen miteinander verglichen werden. Der Mittelwert darf dann natürlich nicht Null sein.

Beispiel:
Die Stichprobenwerte sind 3;4;5; $\bar{x} = 4;\ \ s_x = 0{,}82$

Stellen die Werte Meter dar, und man rechnet sie in Zentimeter um, so ist $\bar{x} = 400;\ \ s_x = 82$ bei gleich aussehender Verteilung. Der Variationskoeffizient ist jedoch in beiden Fällen gleich.

vereinfachende Formel für die Varianz

Die Differenzen sind zumeist keine glatten Zahlen, die Quadrate haben doppelt so viele Nachkommastellen. Dies kann man durch eine vereinfachende Formel umgehen.

Abkürzend verzichten wir hier auf den Faktor $\frac{1}{n-1}$ und rechnen

ohne Klassierung, dann ohne $\frac{1}{n-1}$ aber mit Klassierung, und noch mit $\frac{1}{n-1}$ und mit Klassierung.

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$
$$= \sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2$$
$$= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$
$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{xx} = \sum_{i=1}^{k}(x_i - \bar{x})^2 n_i$$
$$= \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)n_i$$
$$= \sum x_i^2 n_i - \sum 2x_i\bar{x}n_i + \sum \bar{x}^2 n_i$$
$$= \sum x_i^2 n_i - 2n\bar{x}^2 + n\bar{x}^2$$
$$= \sum_{i=1}^{k} x_i^2 n_i - n\bar{x}^2$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i$$
$$= \frac{1}{n-1}\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)n_i$$
$$= \frac{1}{n-1}(\sum x_i^2 n_i - \sum 2x_i\bar{x}n_i + \sum \bar{x}^2 n_i)$$
$$= \frac{1}{n-1}(\sum x_i^2 n_i - 2n\bar{x}^2 + n\bar{x}^2)$$
$$= \frac{1}{n-1}(\sum_{i=1}^{k} x_i^2 n_i - n\bar{x}^2)$$

Streuungsintervalle

Als Maß für die Form und die Breite der Verteilung dient die Standardabweichung. Ausgehend vom Mittelwert legen wir äquidistante **Streuungsintervalle** fest und bestimmen die Anteile der Verteilung in diesen Bereichen.

S_x tragen wir vom Mittelwert ausgehend zu beiden Seiten dreimal ab. Dann haben wir (fast) alle Merkmalwerte abgedeckt.

x_i	n_i	$x_i n_i$
1	2	2
2	5	10
3	8	24
4	7	28
5	2	10
6	1	6
Summe	**25**	**80**

$$\bar{x} = 3{,}2$$

x_i	n_i	$x_i n_i$
1	0	0
2	4	8
3	12	36
4	9	36
5	0	0
6	0	0
Summe	**25**	**80**

Linke Tabelle: Sx = 1.225 Rechte Tabelle: Sx = 0.707

Sx tragen wir vom Mittelwert 3,2 ausgehend zu beiden Seiten dreimal ab.

[1.975; 4.425] 20 von 25 = 80% [2.492; 3.907] 12 von 25 = 48%
[0.751; 5.649] 24 von 25 = 96% [1.786; 4.614] 25 von 25 = 100%
[-0.474; 6.874] 25 von 25 = 100% [1.079; 5.321] 25 von 25 = 100%

Wir stellen fest, dass unabhängig von der Größe der Standardabweichung die Intervalle relativ ähnliche Prozentsätze enthalten. Die Histogramme hatten alle die Form einer Glockenkurve.
Im 3-fachen Streuungsintervall liegen nahezu 100% aller Werte; im 2-fachen 95%, im 1-fachen Streuungsintervall liegen 50% bis 80%.

Wir wählen die Standardabweichung als neue Einheit. Damit wird jede Stichprobe dimensionslos, und die Stichproben sind untereinander leichter zu vergleichen.

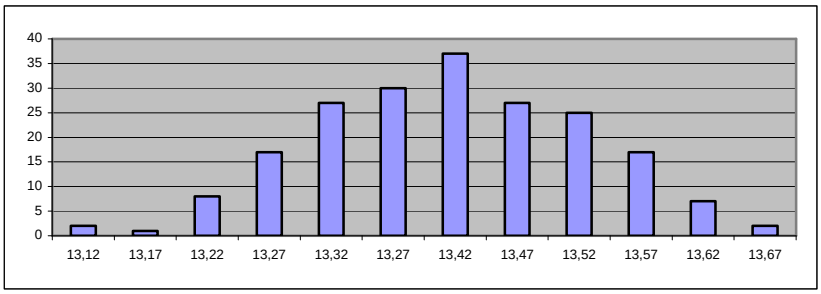
Betrachten wir noch ein letztes Beispiel, das auf einer umfangreicheren Stichprobe basiert.

Die Prüfung von 200 Nietkopfdurchmessern ergab folgende Werte
(12 Klassen, x_i sind die Klassenmitten)

x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$	x_i-mitte
13,12	2	26,24	172,1344	344,2688	-0,28
13,17	1	13,17	173,4489	173,4489	-0,23
13,22	8	105,76	174,7684	1398,1472	-0,18
13,27	17	225,59	176,0929	2993,5793	-0,13
13,32	27	359,64	177,4224	4790,4048	-0,08
13,27	30	398,1	176,0929	5282,787	-0,13
13,42	37	496,54	180,0964	6663,5668	0,02
13,47	27	363,69	181,4409	4898,9043	0,07
13,52	25	338	182,7904	4569,76	0,12
13,57	17	230,69	184,1449	3130,4633	0,17
13,62	7	95,34	185,5044	1298,5308	0,22
13,67	2	27,34	186,8689	373,7378	0,27
Summen	**200**	**2680,1**	**2150,8058**	**35917,599**	

Mittelwert ist 13.40, Standardabweichung 0.121
[-0.121; 0.121] 146 von 200 = 73%
[-0.242; 0.242] 196 von 200 = 98%
[-0.363; 0.363] 200 von 200 = 100%

Wenn wir später die Normalverteilung untersuchen, werden wir feststellen, dass die Anteile in den Streuungsintervallen 68%, 95%, 99% betragen. Man kann mit den Streuungsintervallen einschätzen, ob eine Verteilung eine der Normalverteilung ähnliche Form besitzt.



„Normierung“ $\frac{x_i - \bar{x}}{s_x}$

Wir transformieren die Stichprobenwerte mit $z_i = \frac{x_i - \bar{x}}{s_x}$

und haben so die Verteilung um die y-Achse gruppiert und die x-Achse (besser jetzt z-Achse) mit der neuen Einheit s_x eingeteilt.

Wir finden $\bar{z} = 0 \ und \ s_z = 1$.

Anmerkung:

Diese „Normierung“ mit Hilfe von s_x wird uns später einen weiteren Zugang zur Berechnung des Korrelationskoeffizienten liefern.

Bei der Auswertung von Stichproben bringt die Normierung keinen direkten Vorteil. Da Mittelwert 0 und Varianz 1 dann für alle Verteilungen gelten, können wir sie so nicht mehr unterscheiden. Auch den Variationskoeffizienten können wir nicht mehr berechnen. Allerdings ergeben sich mit den Normierungen Vergleichsmöglichkeiten mit Standard-Verteilungen wie zum Beispiel der Normalverteilung. Bei der Untersuchung der Streuungsintervalle haben wir darauf verwiesen.

4. Beispiele, Übungen

Alle Aufgaben werden mit der Fragestellung:

arithmetisches Mittel, Standardabweichung, Anteile in Streuungsintervallen, Normierung, Vergleich

bearbeitet.

Die Werte findet man im Anhang.

Widerstandsmessung (ohne Klasseneinteilung)

Anzahl Samen in einer Frucht (ohne Klasseneinteilung)

Wachstum von Kiefern (bereits klassiert)

Brenndauer von Glühlampen (mit Klasseneinteilung)

Kupfergehalt (mit Klasseneinteilung)

5. Theorie I: Rechnen mit formalen Summen (lineare Transformationen)

Wir multiplizieren die Stichprobenwerte mit einem Faktor und addieren einen Summanden dazu.

$z_i = ax_i + b$	Die Variable z wird ersetzt.
$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b)$	Die Terme werden getrennt summiert.
$= \frac{1}{n}\sum_{i=1}^{n} ax_i + \frac{1}{n}\sum_{i=1}^{n} b$	a ist konstant, es wird ausgeklammert; Summe über n Summanden *b* ist *nb*.
$= \frac{1}{n} a\sum_{i=1}^{n} x_i + \frac{1}{n} nb$	
$= a\bar{x} + b$	

$z_i = ax_i + b$	Wir transformieren wie oben.
$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$	Das ist die Varianz der z-Werte.
$= \frac{1}{n-1}\sum_{i=1}^{n}((ax_i + b) - (a\bar{x} + b))^2$	z und Mittelwert werden ersetzt.
$= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i + b - a\bar{x} - b)^2$	b fällt weg.
$= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i - a\bar{x})^2$	a wir ausgeklammert; Achtung: quadratisch!
$= \frac{1}{n-1} a^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$	Da steht die Varianz der x-Werte mit Faktor a^2.
$= a^2 s_x^2$	

Das arithmetische Mittel wird genauso transformiert wie die Stichprobenwerte.
Die Varianz wird von additiven Konstanten nicht beeinflusst, multiplikative Konstanten gehen quadratisch in die Varianz ein.
Daraus folgt sofort:
Die um den Mittelwert verminderten Merkmalswerte haben den Mittelwert 0.

$$z_i = x_i - \bar{x} \quad hier \;\; a = 1;\; b = -\bar{x}$$

$$\Rightarrow \bar{z} = 0$$

Die standardisierten Merkmalswerte haben den Mittelwert 0 und die Varianz 1.
Das können wir jetzt bestätigen.

$$z_i = \frac{x_i - \bar{x}}{s_x} = \frac{x_i}{s_x} - \frac{\bar{x}}{s_x} \quad hier \;\; a = \frac{1}{s_x};\; b = -\frac{\bar{x}}{s_x}$$

$$\Rightarrow s_z^2 = 1$$

6. Theorie II: Summe quadrierter Abweichungen ist minimal, wenn sie auf $\bar{x}$ bezogen wird.

Dieser Abschnitt ist optional. Wird aber empfohlen, damit die Schüler die grundlegende Idee auf zweidimensionale Stichproben übertragen können.
Erste Möglichkeit des Nachweises:

$\sum_{i=1}^{n}(x_i - c)^2$	Wir beziehen die Abweichung auf irgendein c.
$= \sum x_i^2 - 2c\underbrace{\sum x_i}_{=n\bar{x}} + \underbrace{\sum c^2}_{=nc^2}$	Binom wird aufgelöst und einzeln summiert.
$= \sum x_i^2 - 2cn\bar{x} \quad + nc^2$	Der erste Summand sieht aus wie Sxx, es fehlt nur $-n\bar{x}^2$
$= \underbrace{\sum x_i^2 - n\bar{x}^2}_{=Sxx} + n\bar{x}^2 - 2cn\bar{x} + nc^2$	Wir ergänzen
$= S_{xx} + n(\bar{x}^2 - 2c\bar{x} + c^2)$	und fassen das Binom zusammen.
$= S_{xx} + n(\bar{x} - c)^2 \geq S_{xx}$	Der zweite Summand ist ≥ 0. Die Summe ist minimal für $c = \bar{x}$, da dann Gleichheit gilt.

Zweite Möglichkeit des Nachweises:

$\sum_{i=1}^{n}(x_i - c)^2$	Wir beziehen die Abweichung auf irgendein c,
$= \sum (x_i - \bar{x} + \bar{x} - c)^2$	ergänzen innerhalb, um auf Sxx zu kommen,
$= \sum ((x_i - \bar{x}) + (\bar{x} - c))^2$	lösen die Klammer auf,
$= \sum (x_i - \bar{x})^2 + \sum 2(x_i - \bar{x})(\bar{x} - c) + \sum (\bar{x} - c)^2$	summieren einzeln.
$= \sum (x_i - \bar{x})^2 + 2(\bar{x} - c)\underbrace{\sum (x_i - \bar{x})}_{=0} + \sum (\bar{x} - c)^2$	Die Abweichungen bezüglich Mittelwert heben sich auf.
$= S_{xx} \quad + 0 \quad + n(\bar{x} - c)^2 \geq S_{xx}$	

Dritte Möglichkeit des Nachweises:
Man kann auch den Term als Quadratfunktion auffassen und das Minimum durch Ableiten bestimmen.

$Abweich(c) = \sum_{i=1}^{n}(x_i - c)^2$	Die Funktionsvariable ist c,
$= \sum x_i^2 - 2c\sum x_i + \sum c^2$	sie ist bezogen auf den Summationsindex konstant
$= \sum x_i^2 - 2cn\bar{x} + nc^2$	Beim Ableiten ist alles Konstante, was nicht c heißt.
$\Rightarrow Abweich'(c) = 0 - 2n\bar{x} + 2nc$	Wegen oben offener Parabel liegt ein Minimum vor.
$= 2nc - 2n\bar{x}$	
$= 2n(c - \bar{x}) = 0 \quad falls \quad c = \bar{x}$	Man kann also sagen, dass das arithmetische Mittel die Stichprobe insofern optimal repräsentiert, als die quadrierten Abweichungen auf dieses Mittel bezogen minimal sind.

Zweidimensionale Stichproben

1. Aufgabe der zweidimensionalen Statistik und grundlegende Begriffe

Bisher haben wir Stichproben untersucht, deren Elemente durch ein einzelnes Merkmal ausgezeichnet waren: Größe, Anzahl, Preis etc.
Jetzt wollen wir Stichproben untersuchen, deren Elemente aus Paaren bestehen: Alter/Preis, Länge/Gewicht, Stromstärke/Spannung etc.
Wir fragen, ob (1) eine gegenseitige Abhängigkeit der Messwerte zu vermuten ist, und dann, ob man (2) die vermutete Abhängigkeit rechnerisch beschreiben kann, und abschließend, (3) wie gut die rechnerische Beschreibung ist.
Erste Erkenntnisse erhalten wir durch eine graphische Darstellung der Paare im Koordinatensystem, die **Punktwolke**. Die Werte scheinen mehr oder weniger stark voneinander abzuhängen, wir sagen: zu **korrelieren**.
Entscheiden wir uns der Einfachheit halber für eine vermutete lineare Abhängigkeit, dann versuchen wir, eine Gerade zu finden, welche die Punktwolke repräsentiert. Wir nennen sie **Regressionsgerade**. Die Qualität der Repräsentanz endlich beschreiben wir mit dem **Korrelationskoeffizienten**.

2. Einstieg

Es gibt verschiedene sinnvolle und motivierende Einstiegsaufgaben.
Biologisch orientiert: Zirpen von Grillen und Außentemperatur
Wirtschaftlich: Alter und Preis von Kfz
Gibt es einen funktionalen Zusammenhang in diesen Stichproben?
Eine vorliegende Datensammlung wird zunächst oberflächlich untersucht. Eine Punktwolke wird gezeichnet und Vermutungen werden geäußert.

3. rechnerische Lösung

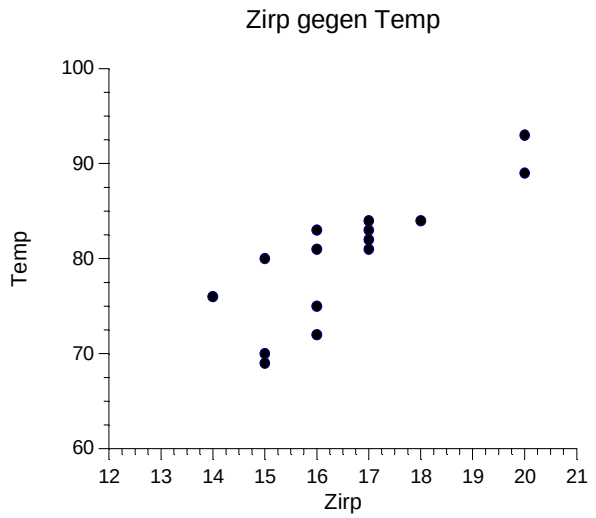
Grillen zirpen

Wenn Grillen die Flügel aneinander reiben, ertönt das charakteristische Zirpen. Wissenschaftler haben festgestellt, dass in Abhängigkeit von der Temperatur Grillen mehr oder weniger schnell zirpen.

Zirp / sec	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Temp	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

Kann man ohne Thermometer mit Hilfe der Grillen die Temperatur feststellen?
(Dieses Beispiel ist insofern interessant, als die Temperatur in Fahrenheit angegeben ist. Dies führt zu der Frage, ob und wie Transformationen die Kenngrößen beeinflussen.)

Wir zeichnen eine Punktwolke.



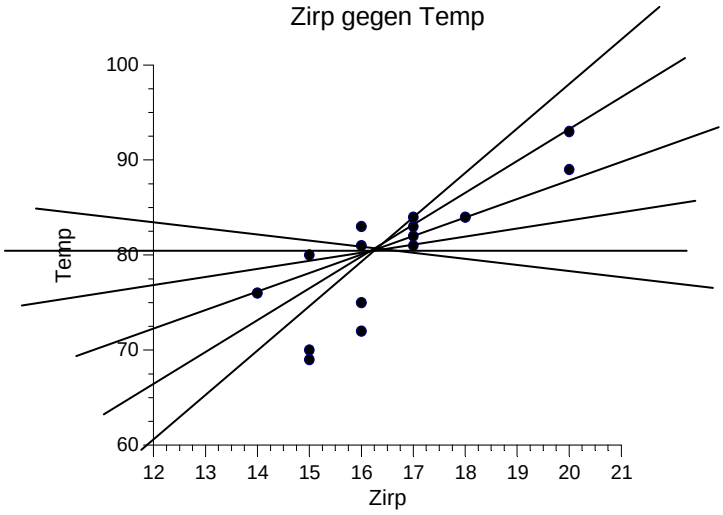
Die Lage der Wolke deutet auf einen Zusammenhang hin.

Zusammenhang der Größen beschreiben (einfachstes Modell: linear)

Je schneller das Zirpen, desto höher die Temperatur. Wir entscheiden uns zunächst für den Ansatz des linearen Modells. Das bedeutet, dass wir versuchen, die Punktwolke durch eine Lineare Funktion zu repräsentieren.

Gerade anpassen (Schwerpunkt der Wolke auf der Geraden)

Wir legen verschiedene Geraden durch die Wolke und verwerfen offensichtlich schlechte Lagen.
Wenn die Punktwolke eine „Richtung“ hat, sollte diese durch die Steigung der Geraden wiedergegeben werden. Eine Best-Gerade sollte durch die „Mitte“ verlaufen. Die Mitte könnte der Punkt der Mittelwerte sein, wir nennen ihn **Schwerpunkt** der Wolke. Wir müssen also die optimale Steigung bestimmen.



Die Gerade soll ein Repräsentant für die Punktwolke sein.

Unter den vielen Möglichkeiten, das Streuungsmaß einer Punktwolke bzgl. einer gegebenen Geraden festzulegen, hat sich die quadrierte <u>vertikale</u> Abweichung als Standard etabliert. Die hierdurch definierte Gerade wird als „die Regressionsgerade“ bezeichnet (entspricht z.B. in Excel der sog. Trendlinie).

Damit ist klar, dass der Schwerpunkt der Wolke $S(\overline{x}/\overline{y})$ auf dieser Geraden liegen muss, damit dessen Abweichung Null wird.

Die Schüler werden sich daran erinnern, was optimale Repräsentanz bei eindimensionalen Stichproben bedeutet. Das arithmetische Mittel ist insofern optimal, als die Summe der quadrierten Abweichungen minimal ist.

Im Zweidimensionalen spielt die Gerade die Rolle des arithmetischen Mittels. So verlangen wir entsprechend minimale quadrierte vertikale Abweichungen der Punkte auf der Geraden zu den Punkten der Wolke!

Abweichungen minimieren (Parabel-Tiefpunkt)

Die Regressionsgerade sei $g(x) = ax + b$

Da die Gerade durch den Schwerpunkt verläuft, können wir den Parameter b ersetzen:

$$g(\overline{x}) = a\overline{x} + b = \overline{y} \Rightarrow b = \overline{y} - a\overline{x}$$

$$\Rightarrow g(x) = ax + \overline{y} - a\overline{x}$$

Wir betrachten nun erneut die Funktion *v(m)* und bestimmen deren Minimum.
Zur Erinnerung: Die y_i sind die realen y-Werte der Messpunkte (x_i / y_i), während $g(x_i)$ die zugehörigen y-Werte auf der Geraden darstellen.

$$v(a) =$$
$$= \sum_{i=1}^{n} (y_i - g(x_i))^2 \qquad | \, g(x) \;\; ersetzen$$
$$= \sum ((y_i - (ax_i + \overline{y} - a\overline{x}))^2 \qquad | \, nach\ y\ und\ x\ sortieren$$
$$= \sum ((y_i - \overline{y}) - a(x_i - \overline{x}))^2 \qquad | \, Binom\ \ auflösen\ \ und\ \ getrennt\ \ Summieren$$
$$= \sum ((y_i - \overline{y})^2 - 2a(y_i - \overline{y})(x_i - \overline{x}) + a^2 (x_i - \overline{x})^2)$$
$$= \sum (y_i - \overline{y})^2 - 2a \sum (y_i - \overline{y})(x_i - \overline{x}) + a^2 \sum (x_i - \overline{x})^2$$

Wir benutzen weiter die abkürzenden Schreibweisen

$$Sxx = \sum_{i=1}^{n} (x_i - \overline{x})^2$$
$$Syy = \sum_{i=1}^{n} (y_i - \overline{y})^2$$
$$Sxy = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

und erhalten somit

$$v(m) = \qquad Syy \qquad - 2aSxy \qquad\qquad + a^2 Sxx$$

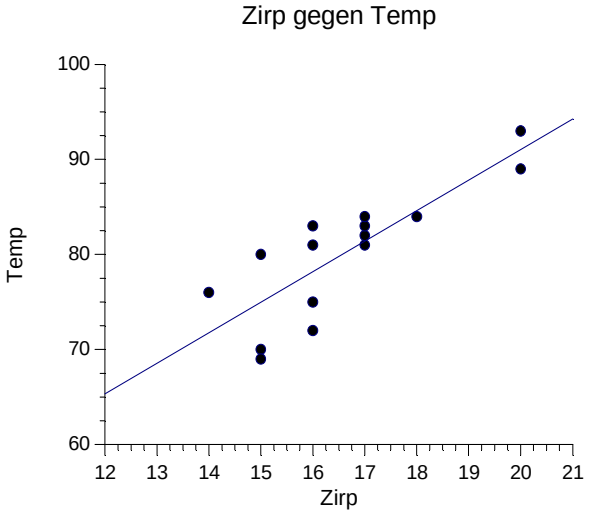
v(a) soll minimiert werden. Wovon ist *v* abhängig? Natürlich von den einzelnen Abweichungssummen *Sxx*, *Syy*, *Sxy*. Aber das sind Konstante, wenn wir die Tabelle einmal ausgewertet haben. Also bleibt die Abhängigkeit von *m*. Dann ist *v(m)* aber eine Quadratfunktion.

$$v(a) = a^2 Sxx - 2aSxy + Syy$$
$$= Sxx(a^2 - 2a\frac{Sxy}{Sxx} + \frac{Syy}{Sxx})$$

Kenntnis der Normalparabel oder Ableitung zeigt: Das Minimum liegt bei $a_{\text{opt}} = \frac{Sxy}{Sxx}$.

Wir finden also $g(x) = a_{opt} \cdot (x - \overline{x}) + \overline{y} = \frac{Sxy}{Sxx}(x - \overline{x}) + \overline{y}$

So können wir jetzt die optimale Gerade in die Punktwolke einzeichnen.



4. Qualitätsmaß für die Regressionsbeziehung (relative Abweichungssumme)

Ohne die Regressionsbeziehung, also die kalkulierte Abhängigkeit der y-Werte von den x-Werten (hier die Temperaturen von der Zirpzahl) wäre allein das arithmetische Mittel $\bar{y}$ der Repräsentant der y-Werte gewesen. Daran gemessen hatten diese Werte eine gewisse Varianz.
Dem arithmetischen Mittel entspricht jetzt in der zweidimensionalen Probe die Horizontale durch den Schwerpunkt.
Da wir aber die Abhängigkeit von den x-Werten ins Spiel bringen, mussten wir diese Horizontale in die „bessere“ Lage drehen. Der Varianz im Eindimensionalen entspricht jetzt die mittlere Summe der quadrierten vertikalen Abweichungen von dieser „gedrehten“ Geraden, d. h. der Regressionsgeraden.

Der Einfachheit halber rechnen wir weiter mit den Summen, die nicht mit dem Stichprobenumfang gemittelt wurden.
„Die vertikale Abweichung“ der Punkte von der Regressionsgeraden vergleichen wir mit ihrer „vertikalen Abweichung“ vom Mittelwert ihrer y–Werte.

Dazu bestimmen wir zunächst $v(a_{opt})$ und dann setzen diesen Wert in Relation zu Syy.

$$v(a_{\text{opt}}) = \sum (y_i - g(x_i))^2$$
$$= \sum ((y_i - \overline{y}) - a_{\text{opt}}(x_i - \overline{x}))^2$$
$$= Syy - 2a_{\text{opt}}Sxy + a_{\text{opt}}{}^2 Sxx$$
$$= Syy - \frac{2(Sxy)^2}{Sxx} + (\frac{Sxy}{Sxx})^2 \cdot Sxx$$
$$= Syy - \frac{(Sxy)^2}{Sxx}$$

$v(a_{opt})$ ist also offensichtlich kleiner als Syy.

D.h. die vertikale Abweichung der Punkte von der Regressionsgeraden ist kleiner als die vertikale Abweichung vom Mittelwert ihrer y–Werte.

Man betrachtet den Subtrahenden $\frac{(Sxy)^2}{Sxx}$ als den Anteil, der durch die Regression „erklärt“ wird.
Dieser Anteil wird bezogen auf Syy, also zu einem relativen Anteil gemacht. Um welchen relativen Anteil wird Syy verkleinert, wenn man die Beziehung von y zu x einrechnet? Das ist offensichtlich $\frac{(Sxy)^2}{Sxx \cdot Syy}$.

Der Term ist ein Bestimmtheitsmaß für die Beziehung zwischen x und y.

$\mathbf{B} = \frac{(Sxy)^2}{Sxx \cdot Syy}$ liegt zwischen 0 und 1, je näher bei 1, desto stärker ist die Beziehung.

Ist B=0, dann gilt $v(a_{opt}) = S_{yy}$, und offenbar hat sich durch Einbeziehen der x-Werte nichts verändert.
Die y-Werte sind von x unabhängig.
Ist B=1, dann ist $v(a_{opt}) = 0$ und die Punkte liegen alle auf der Geraden.

Man definiert den Korrelationskoeffizienten

$$r_{xy} = \sqrt{B} = \sqrt{\frac{(Sxy)^2}{Sxx \cdot Syy}} = \frac{Sxy}{\sqrt{Sxx \cdot Syy}}$$

mit dem Vorzeichen von Sxy.
Das Vorzeichen liefert den Trend der Beziehung, man spricht daher von positiver oder negativer Korrelation.
Die Frage, wie groß r sein muss, um die funktionale Abhängigkeit ziemlich gut zu erfassen, kann hier nicht beantwortet werden. Man sollte aber beachten, dass zum Beispiel bei einem Korrelationskoeffizienten von r=0.7 das Bestimmtheitsmaß nur B=0.49 beträgt, was die „Größe“ von r etwas relativiert.

5. Theorie III: Steigung und Korrelationskoeffizient unter linearen Transformationen

Wir haben bei der Untersuchung eindimensionaler Stichproben den Mittelwert und die Varianz linearen Transformationen unterworfen.
In den Beispielen Zirpen, Nietköpfe und Körpergröße wäre eine Änderung des Maßstabes denkbar: Temperatur in Celsius statt Fahrenheit, Durchmesser in Zoll statt Millimeter, Körpergröße in Fuß statt Zentimeter.
Man könnte jetzt untersuchen, wie sich die Parameter der Regressionsgeraden und der Korrelationskoeffizient unter linearer Transformation verhalten. Hier wäre speziell die Frage interessant, wie sich die Normierung auswirkt.
Die x-Werte werden zu z-Werten, die y-Werte werden zu w-Werten transformiert.

$$z_i = \frac{(x_i - \overline{x})}{s_x} \quad \wedge \quad w_i = \frac{(y_i - \overline{y})}{s_y}$$

$$\Rightarrow \overline{z} = \frac{(\overline{x} - \overline{x})}{s_x} = 0 \quad \wedge \quad \overline{w} = 0$$

$$\Rightarrow s_z^2 = \frac{s_x^2}{s_x^2} = 1 \quad \wedge \quad s_w^2 = 1$$

$$s_x^2 = \frac{1}{n} S_{xx} \wedge \cdots \Rightarrow S_{xx} = n \cdot s_x^2 \wedge \cdots \Rightarrow \sqrt{S_{xx}} = \sqrt{n} \cdot s_x \wedge \cdots$$

Die Regressionsgerade verlief durch den Schwerpunkt; dieser liegt jetzt im Ursprung des Koordinatensystems.
Die normierte Regressionsgerade ist also eine Ursprungsgerade.
Welche Steigung hat sie?

$$a_{norm} = \frac{S_{zw}}{S_{zz}}$$

Die Steigung a haben wir allgemein oben hergeleitet,
jetzt ersetzen wir sinngemäß die normierten Größen mit z und w.

$$= \frac{\frac{1}{s_x} \cdot \frac{1}{s_y} \cdot S_{xy}}{\frac{1}{s_x^2} \cdot S_{xx}}$$

Konstante Faktoren bleiben bei der Transformation erhalten,
das sind hier $\frac{1}{s_x}$ und $\frac{1}{s_y}$

$$= \frac{s_x^2 \cdot S_{xy}}{s_x s_y \cdot S_{xx}}$$

Doppelbrüche einrichten

$$= \frac{\frac{1}{n-1} S_{xx} \cdot S_{xy}}{s_x s_y \cdot S_{xx}}$$

s_x^2 ist die mittlere Abweichung.
s_x die Wurzel daraus;

$$= \frac{\frac{1}{n-1} S_{xx} \cdot S_{xy}}{\sqrt{\frac{1}{n-1} S_{xx}} \cdot \sqrt{\frac{1}{n-1} S_{yy}} \cdot S_{xx}}$$

wir kürzen

$$= \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = r_{xy}$$

und erkennen den Korrelationskoeffizienten wieder.

$$r_{zw} = \frac{S_{zw}}{\sqrt{S_{zz} \cdot S_{ww}}}$$

Jetzt bestimmen wir noch den Korrelationskoeffizienten der normierten Werte.

$$= \frac{\frac{1}{s_x s_y} S_{xy}}{\sqrt{\frac{1}{s_x^2} S_{xx} \cdot \frac{1}{s_y^2} S_{yy}}}$$

Die Tricks sind dieselben wie oben.

$$= \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = r_{xy}$$

Und wir erkennen, dass der Korrelationskoeffizient beim Normieren gleichbleibt.

Die oben benutzten Normierungen führen also dazu, dass die Steigung gleich dem Korrelationskoeffizienten wird; sie beeinflussen den Korrelationskoeffizienten selbst nicht.

6. Beispiele, Übungen

Körpergröße und Gewicht

bei Kindern

bei Jugendlichen

„alles wird teurer“: Preisentwicklung

„Mathe eins, Sport fünf“: Noten in Mathe und Sport

„Fernsehen macht träge“: Fernsehkonsum und Zeiten sportlicher Aktivität

Kfz, Alter-Preis; Doppelwurf; Wassergehalt (siehe Anhang)

Anhang (Aufgaben)

1. Widerstandsmessung

Ohm	36,1	38,0	37,2	36,3	37,5	37,4	37,1	37,1	36,2	37,9

2. Samen pro Frucht einer Pflanze

Anzahl Samen x_i	3	4	5	6	7	8	9	10	11
Anzahl Früchte n_i	1	2	8	13	22	45	63	23	1

3. Kiefernhöhe in cm

x_i	60	80	100	120	140	160	180	200	220	240	260
n_i	1	1	2	9	14	22	31	27	9	6	3

4. Brenndauer in Stunden

Klassenmitte	550	650	750	850	950	1050	1150	1250	1350	1450	1550
n_i	3	4	6	13	20	19	13	8	1	2	1

5. Kupfergehalt

Urliste 33 Werte, klassieren, erste Klasse 75,20..75,30

75,39 76,09 75,80 76,01 75,21 75,38 75,70 75,56
75,40 75,38 75,80 75,56 75,36 76,04 75,84 75,38
75,60 76,09 75,72 76,02 75,58 76,08 75,54 75,30
75,36 76,04 75,40 75,68 75,90 75,38 75,34 75,24 75,68

6. Kfz:Alter-Preis

x in Jahren, y in Euro

x	y
1	7180
1	6740
2	5580
2	5200
3	4950
4	2760
4	3340
6	2000
7	2200
10	600

7. Wurf zweier Würfel

x	6	4	2	4	3	6	1	2	4
y	6	5	3	2	5	4	2	1	3

8. Gewicht und Wassergehalt von pflanzlichem Gewebe

x und y in Gramm

x	37	32	33	50	39	52	78	91	101	99	101	125
y	13,2	13,4	12,8	11,8	13,2	13,5	14,2	14,7	14,5	18,4	18,4	18,7