1 Grundlagen statistischer Versuchsauswertung

1.1 Statistische Daten

Schon vor der Erhebung von Daten sollten erste statistische Überlegungen mit in das Erhebungsprogramm aufgenommen werden. Primäres Ziel sollte hierbei die Repräsentativität sein, sodass im Vorfeld genau überlegt werden sollte, wie viel Parallelen zum Beispiel pro Messparameterkombination gewählt oder wie viele Personen befragt werden sollten. Wie Abbildung 1 darstellt beginnt jede Datenerhebung mit einer sorgfältigen Planung, auf welche die Erhebung aufbaut. Die erhobenen Daten sollten darauffolgend mithilfe der deskriptiven Statistik aufgearbeitet und ausgewertet werden. An einigen Stellen können auch Auswertungsmethoden der induktiven Statistik vonnöten sein. Nicht zuletzt sollte auf Grundlage der statistischen Auswertung eine Interpretation vorgenommen werden, mit welcher Rückschlüsse auf die Zielsetzung gezogen werden können.

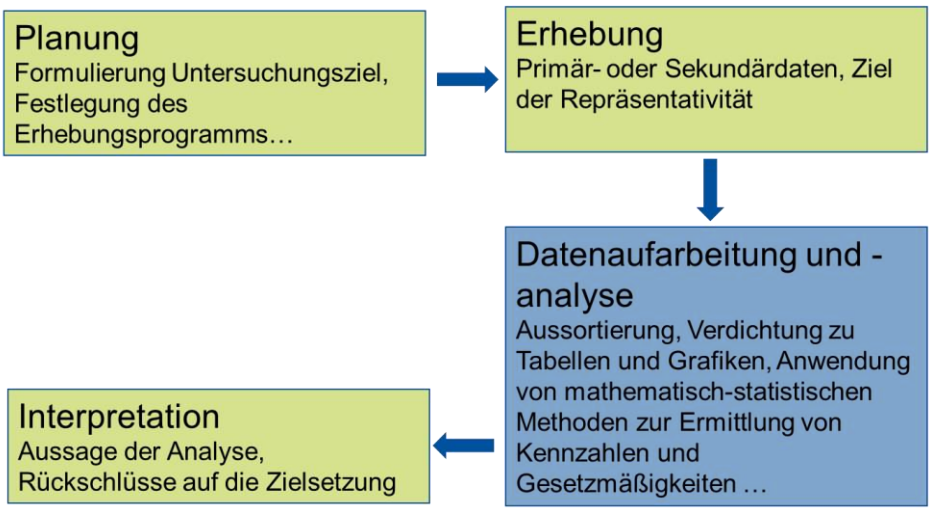


Abbildung 1: Schema der Arbeitsschritte zur Erhebung und Auswertung von Messdaten

Mithilfe von statistischen Daten soll eine allgemeine Aussage über eine Grundgesamtheit getroffen werden. Als Beispiel dient in Abbildung 2 die Forderung nach einer allgemeinen Aussage über die Massenabnahme von thermisch modifizierten Hölzern, welche hier die Grundgesamtheit darstellt. Da es nicht mög-

lich ist, eine Grundgesamtheit in Gänze zu testen, wird lediglich eine repräsentative Stichprobe gewählt, welche aus Einzelobjekten, den so genannten Merkmalsträgern besteht. Jene besitzen das Merkmal dessen Merkmalsausprägung bei der Erhebung der Daten ermittelt werden soll.

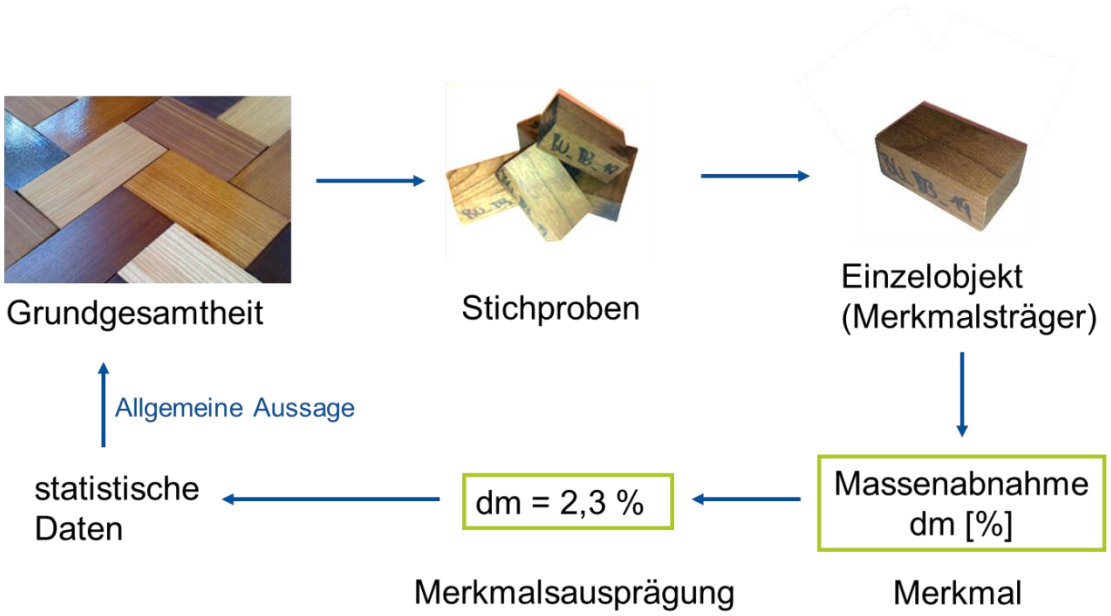


Abbildung 2: Statistische Datenerhebung und Auswertung zur Formulierung einer allgemeinen Aussage über eine Grundgesamtheit am Beispiel der Massenabnahme von thermisch modifizierten Hölzern

Hierbei gilt zu unterscheiden, ob univariante oder multivariante Merkmale vorliegen. Bei Univariate Merkmalen wird nur eine Eigenschaft wird gemessen, sie sind eindimensional:

$x_1, x_2, \ldots, x_n$ wobei n Stichprobenumfang

Multivariate Merkmale sind Kombinationen mehrerer einzelner Merkmale bzw. Merkmale, deren Ausprägungen aus Merkmalsausprägungen mehrerer einzelner Merkmale bestehen, sie sind mehrdimensional:

$x_{i1}, x_{i2}, \ldots, x_{im}$ wobei $i \in \{1, \ldots n\}$

Am Ende der Erhebung ist die Sammlung der Rohdaten abgeschlossen und die Aufbereitung der Daten zu statistischen Daten gefordert. Mithilfe der statistischen Daten lässt sich im Anschluss im günstigen Fall durch Interpretation eine allgemeingültige Aussage über die Grundgesamtheit tätigen.

1.2 Aufbereitung von Daten

Um Daten zu verdichten und ähnliche Merkmalsausprägungen zusammen zu fassen, sollten Merkmale durch Häufigkeiten dargestellt werden.

Eine Stichprobe mit einem Merkmal X besteht aus n Elementen (Merkmalsausprägungen) $x_1, x_2, \ldots, x_n$ der sogenannten Urliste, wobei n den Stichprobenumfang (Anzahl aller Merkmalsaus-prägungen) darstellt. Häufigkeiten beschreiben die Anzahl des Auftretens der Merkmalsausprägungen in der Urliste. Unterschieden werden zwei Arten der Häufigkeit:

Absolute Häufigkeit

Anzahl von Beobachtungswerten, die mit einer bestimmten Merkmalsausprägung identisch sind

→ Summe aller absoluten Häufigkeiten ergibt den Stichprobenumfang n

Relative Häufigkeit

$$\text{relative Häufigkeit} = \frac{\text{absolute Häufigkeit}}{\text{Stichprobenumfang}} \; (\cdot 100\ \%)$$

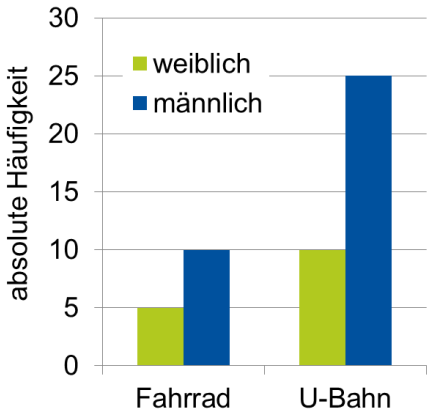
→ Summe aller relativen Häufigkeiten ergibt 1 (oder 100 %)

Um Häufigkeiten von Merkmalen nachvollziehbar darstellen zu können, ist es sinnvoll die aufbereiteten Daten in Tabellen und Diagrammen abzubilden. Hierbei ist es äußert wichtig, dem Leser alle Informationen in den Grafiken zukommen zu lassen, da es sonst zu einer Manipulation der Aussage der Daten kommen kann. Explizit bedeutet dies bei der Darstellung von relativen Häufigkeiten, dass der Stichprobenumfang unbedingt mit angegeben werden muss. Inwieweit unterschiedliche Darstellungen bei gleichen Daten zu verschiedenen Aussagen führen können, verdeutlicht das folgende Beispiel in Verbindung mit Tabelle 1 und Abbildung 3:

Das Ergebnis einer Befragung von Studenten, ob sie eher mit der U-Bahn oder mit dem Fahrrad zur Uni kämen, ergab folgende Häufigkeitstabelle:

Tabelle 1: Beispiel eines Ergebnisses einer Befragung von weiblichen und männlichen Studenten, ob sie eher mit der U-Bahn oder mit dem Fahrrad zur Uni kämen. Darstellung der absoluten und relativen Häufigkeiten

	Fahrrad	U-Bahn	Summe
	absolute Häufigkeiten		
weiblich	5	10	15
männlich	10	25	35
	relative Häufigkeiten		
weiblich	33,3 %	66,7 %	100 %
männlich	28,6 %	71,4 %	100 %



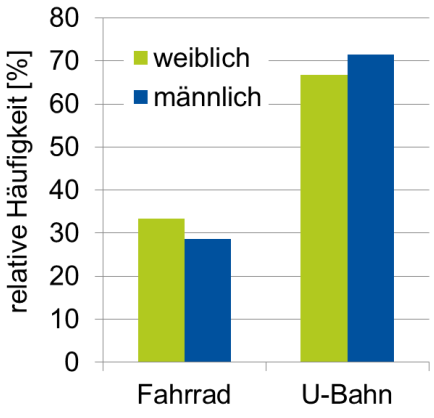


Abbildung 3: Darstellung des Datensatzes aus Tabelle 1 als Balkendiagramme in absoluten und relativen Häufigkeiten

Obwohl beide Darstellungen aus Abbildung 3 auf demselben Datensatz basieren, suggerieren die Darstellung der Häufigkeiten völlig unterschiedliche Aussagen über die geschlechterspezifische Benutzung des Fortbewegungsmittels.

1.3 Statistische Kennwerte

1.3.1 Eindimensionale Stichproben

1.3.1.1 Lagemaße

Bei der Befragung mehrerer Personen zu einem quantitativen Merkmal oder der Ermittlung einer physikalischen Größe mehrerer Prüfkörper interessiert meistens die mittlere Angabe der Größe bzw. des Merkmales.

In der Statistik existiert eine Vielzahl von unterschiedlichen Mittelwerten. Das arithmetische Mittel ist der am häufigsten verwendete Mittelwert. Definiert ist das arithmetische Mittel als der Quotient aus der Summe aller Messwerte und der Anzahl der Messwerte (Gleichung 1).

Gleichung 1: Berechnung des arithmetischen Mittels

Arithmetisches Mittel

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad n = \text{Anzahl der Messwerte}$$

Unbedingt zu beachten beim arithmetischen Mittel ist dessen Anfälligkeit gegenüber Ausreißern. Diese können zu einer erheblichen Manipulation des arithmetischen Mittels führen. Wird beispielsweise bei der Datenerhebung ein Tippfehler gemacht, sodass das Komma fälschlicher Weise eine Dezimalstelle verschoben wird (106,3 g anstatt 10,63 g) wird dies auch bei einer ausreichend großen Anzahl von Messwerten zu einer erheblichen Erhöhung des Wertes für das arithmetische Mittel führen.

Ein weiterer Mittelwert, welcher hingegen robust gegenüber Ausreißern ist, stellt der Median (oder auch Zentralwert) dar. Dieser liegt in der Mitte einer aufsteigend geordneten Liste von Messwerten und teilt diese in die 50 % kleinsten bzw. 50 % größten Messwerte (Gleichung 2).

Gleichung 2: Berechnung des Medians

Median

$$\tilde{X} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & ; \text{falls Anzahl der Messwerte n ungerade} \\ \frac{1}{2} \cdot x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} & ; \text{falls Anzahl der Messwerte n gerade} \end{cases}$$

Neben dem Median, welches auch als 0,5-Quantil bezeichnet wird, existieren weitere Quantile, die eine aufsteigend geordnete Liste von Messdaten in zwei Bereiche aufteilen (Gleichung 3).

Gleichung 3: Berechnung von Quantilen

p-Quantil

$$\tilde{X}_p = \begin{cases} \frac{1}{2} \cdot (x_k + x_{k+1}) & \text{für } k=n \cdot p \text{ ganzzahlig} \\ x_k & \text{für } k=n \cdot p \text{ nicht ganzzahlig, gilt der nächstgrößere ganze Wert} \end{cases}$$

n; Anzahl der Prüfkörper

Die Verwendung diverser Quantile ist dann sinnvoll, wenn nicht die ganze Stichprobe vermessen werden kann, sodass kein wahrer Mittelwert gebildet werden kann und ein Quantil als Vergleichsgröße dient. So kann das 0,25-Quantil gewählt werden, welches die Messreihe in 25 % kleinste und 75 % größte Messwerte aufteilt. Diese auch häufig als Quartil bezeichneten Quantile teilen die Daten in vier gleiche Teile; sodass die $\tilde{X}_{0,25}$, $\tilde{X}_{0,5}$, $\tilde{X}_{0,75}$ Quantile häufig als unteres, mittleres und oberes Quartil bezeichnet werden. Perzentile teilen die Daten in 1 % Segmente, sodass das 97 % Perzentil identisch mit $\tilde{X}_{0,97}$ Quantil ist.

1.3.1.2 Streumaße

Die alleinige Darstellung von Lagemaßen ist statistisch wenig sinnvoll, da die übrigen Messwerte um die Lagemaße wie das des arithmetischen Mittels und des Medians streuen. Diese Streuung muss bestimmt werden, um das Abwei-

chungsverhalten der Messwerte vom arithmetischen Mittel und dem Median zu messen, damit eine verlässliche Aussage der Messwerte gegeben ist. Eine mögliche Ermittlung der Streuung ist die Berechnung der empirischen Varianz, welches die mittlere quadratische Abweichung der Messdaten vom arithmetischen Mittel darstellt. Sie ist definiert als der Quotient aus der Summe aller Abweichungsquadrate und der Anzahl der Messwerte (Gleichung 4).

Gleichung 4: Berechnung der empirischen Varianz

Empirische Varianz

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2$$

n; Anzahl der Prüfkörper

$\overline{X}$; Arithmetisches Mittel

Damit jedoch Lage- und Streumaße dieselbe Einheit aufweisen, hat sich die Angabe der Standardabweichung anstelle der empirischen Varianz etabliert. Die Standardabweichung berechnet sich aus der Quadratwurzel der empirischen Varianz (Gleichung 5).

Gleichung 5: Berechnung der Standardabweichung

Standardabweichung

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2}$$

n; Anzahl der Prüfkörper

$\overline{X}$; Arithmetisches Mittel

In jeder Bachelor- sowie Masterarbeit sollte die Angabe des arithmetischen Mittels immer in Verbindung mit der Standardabweichung erfolgen:

Arithmetisches Mittel ± Standardabweichung; $\overline{X} \pm s$

Zum Beispiel sollte die Aussage über den Feuchtegehalt immer relativ als „Die Feuchte liegt im Mittel bei 9,0 % ± 0,4“ formuliert werden. Auch die grafische Darstellung sollte die oben genannten statistischen Kennwerte aufweisen (Abbildung 1)

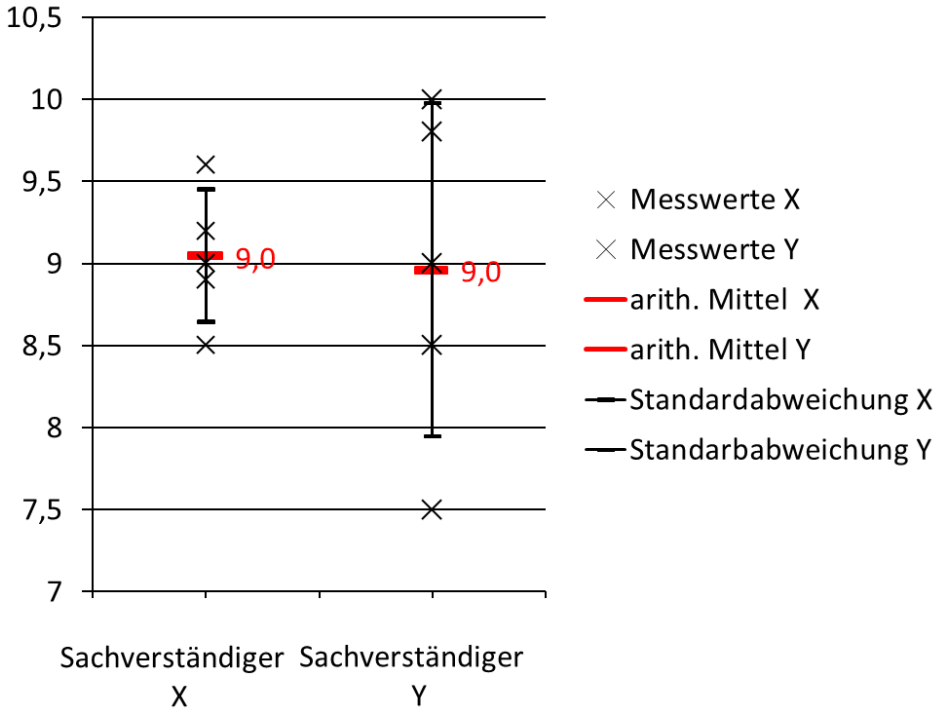


Abbildung 4: Beispiel einer grafischen Darstellung von Messwerten unter Angabe von statistischen Kennwerten wie dem des arithmetischen Mittels und der Standardabweichung

Beim Vergleich der Streuung von zwei Datensätzen, deren gemessene Werte sich hinsichtlich der Größenordnung stark voneinander unterscheiden (z.B. Streuung des Alters von Bäumen und Menschen) ist die alleinige Angabe der Standardabweichung meist wenig sinnvoll. An dieser Stelle eignet sich die statistische Kenngröße des Variationskoeffizienten. Dieser erlaubt sogar den Vergleich von Datensätzen mit unterschiedlichen Einheiten ohne das umgerechnet werden muss, weil der Variationskoeffizient einheitslos bzw. dimensionslos ist. Der Variationskoeffizient kann als normierte Standardabweichung verstanden werden und ist definiert als der Quotient aus empirischer Standardabweichung und arithmetischen Mittel.

Gleichung 6: Berechnung des Variationskoeffizienten

Variationskoeffizient

$$V = \frac{s}{\overline{X}} \cdot 100\%$$

s; empirische Standardabweichung
$\overline{X}$; Arithmetisches Mittel

1.3.1.3 Schätzfehler

Die Abweichungen der Messwerte vom Mittelwert lassen sich durch die sogenannte Normalverteilung entweder exakt oder wenigstens in sehr guter Näherung beschreiben. Die typische Form der Normalverteilung lässt sich durch die Gauß'sche Glockenkurve darstellen. Die Form dieser Glockenkurve ist lediglich von dem arithmetischen Mittel und der Standardabweichung abhängig.

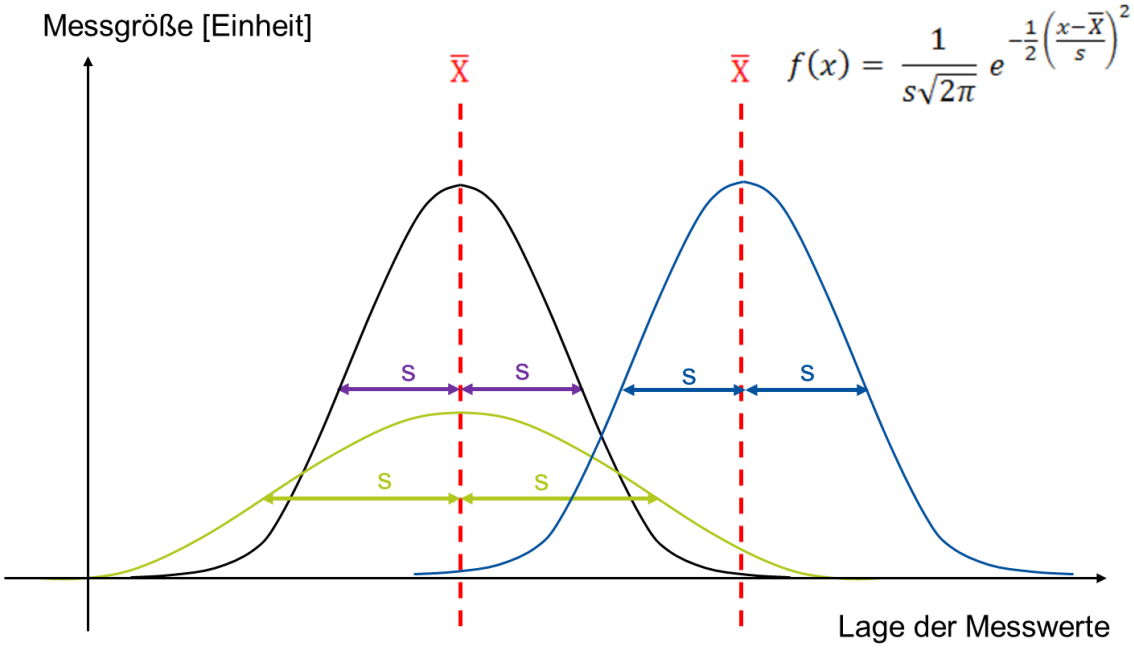


Abbildung 5: Abhängigkeit der Glockenkurve der Normalverteilung vom arithmetischen Mittel und der Standardabweichung

Um auf das „wahre" arithmetische Mittel der zu untersuchenden Grundgesamtheit schließen zu können, wird davon ausgegangen, dass alle Mittelwerte möglicher Stichproben ebenfalls normalverteilt um das „wahre" arithmetische Mittel

der Grundgesamtheit liegen. Hierzu wird in der induktiven Statistik ein Vertrauensbereich angegeben. Vertrauensintervalle (auch Konfidenzintervalle) bestimmen einen Bereich um das errechnete arithmetische Mittel der Stichprobe, in dem das „wahre“ arithmetische Mittel der Grundgesamtheit mit einer zu benennen Wahrscheinlichkeit liegt (Abbildung 6).

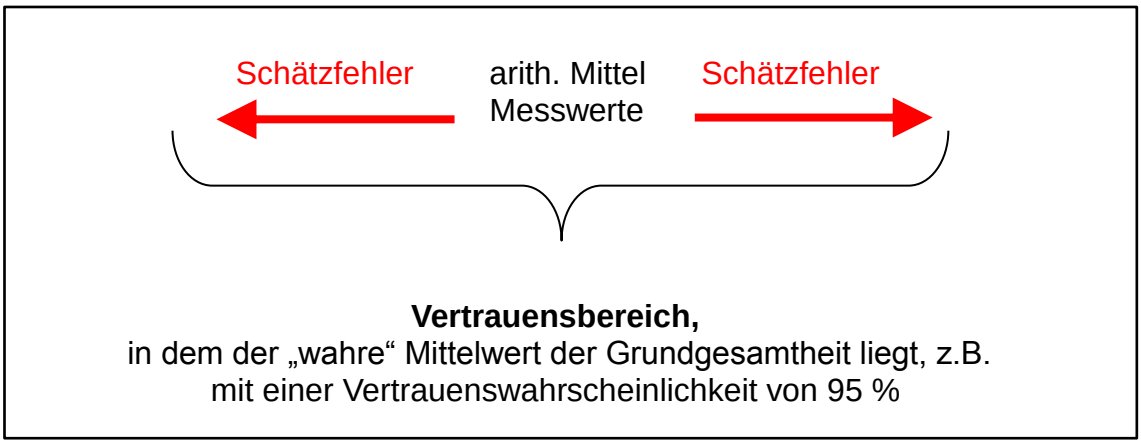


Abbildung 6: Schematische Darstellung des Vertrauensbereiches

Die Berechnung des Schätzfehlers basiert in der am Institut verwendeten Statistik-Auswertung-Excel-Datei auf dem t-Test (Gleichung 7).

Gleichung 7: Ermittlung eines Konfidenzinterfalls

Konfidenzintervall

$$\bar{X} - t_{1-\frac{\alpha}{2},[n-1]} \frac{s}{\sqrt{n}} \quad , \quad \bar{X} + t_{1-\frac{\alpha}{2},[n-1]} \frac{s}{\sqrt{n}}$$

$\bar{X}$; Schätzwert, oft das arithmetische Mittel

α; ist die Wahrscheinlichkeit, dass der Mittelwert der Grundgesamtheit nicht in dem angegebenen Konfidenzintervall liegt → Irrtumswahrscheinlichkeit

1- α; Vertrauenswahrscheinlichkeit

s; Standardabweichung

n; Stichprobenumfang

Grundsätzlich gilt, je kleiner der Vertrauensbereich desto größer ist die Verlässlichkeit der Messdaten einzustufen. Die Größe des Vertrauensbereiches wird maßgeblich von der Streuung und der Anzahl der Messwerte beeinflusst (Abbildung 7). Hier gilt, je größer die Streuung der Messwerte desto größer der

Vertrauensbereich, umgekehrt gilt, je größer die Anzahl der Messwerte desto kleiner der Vertrauensbereich.

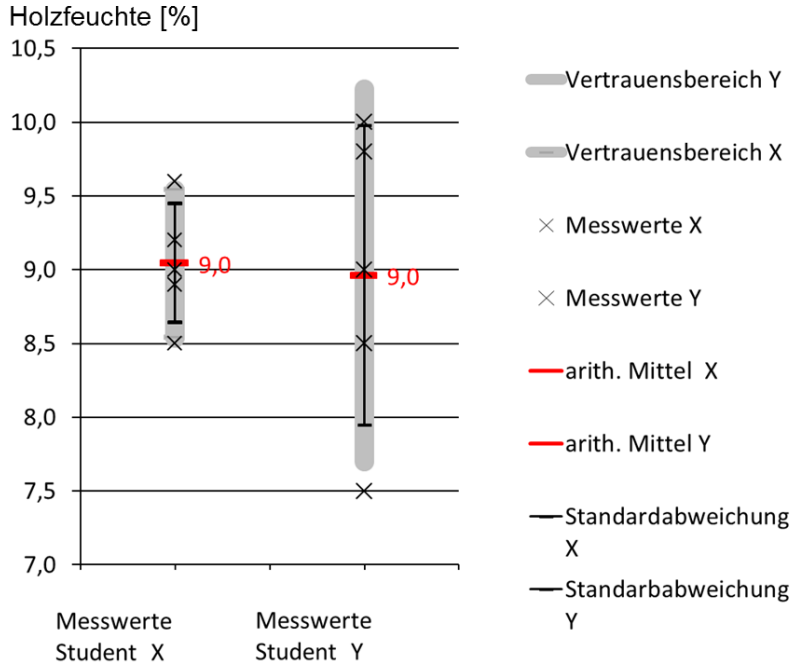


Abbildung 7: Darstellung von Vertrauensbereichen um arithmetische Mittel in Abhängigkeit von der Standardabweichung am Beispiel der Holzfeuchte aus zwei unabhängigen Stichproben (Messwerte Student X oder Y)

1.3.2 Zweidimensionale Stichproben

1.3.2.1 Korrelation

Als Korrelation wird der Zusammenhang zwischen zwei Merkmalen von Stichproben bezeichnet. Dabei werden zwei Merkmale eines Einzelobjektes als Paar von Messwerten (x,y) dargestellt. Jedes Punktepaar wird in Streudiagrammen als Punkt entsprechend seines x und y-Wertes eingetragen. Dadurch kommt es zur Darstellung von Punktwolken (Abbildung 8).

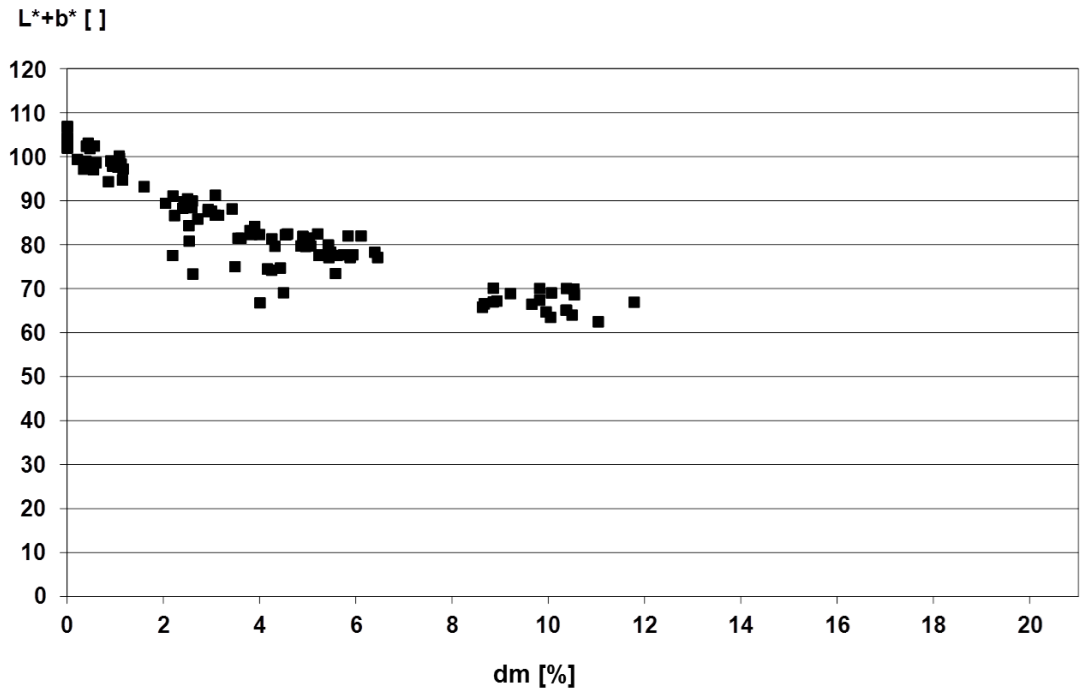


Abbildung 8: Punktwolke aus Messpaaren zweier Merkmale (L*+ b* und dm), eingezeichnet in Koordinatensystem

Bestehen zwischen zwei Merkmalen funktionale Zusammenhänge können diese mithilfe der Methode der kleinsten Fehlerquadrate berechnet werden (Abbildung 9). Dabei wird anhand von Regressionsfunktionen der möglichst minimalste mittlere Abstand zwischen Messwert und durch die Regressionsfunktion geschätzten Wert gesucht, um so die Abhängigkeit zwischen den beiden Merkmalen durch eine Funktion darzustellen.

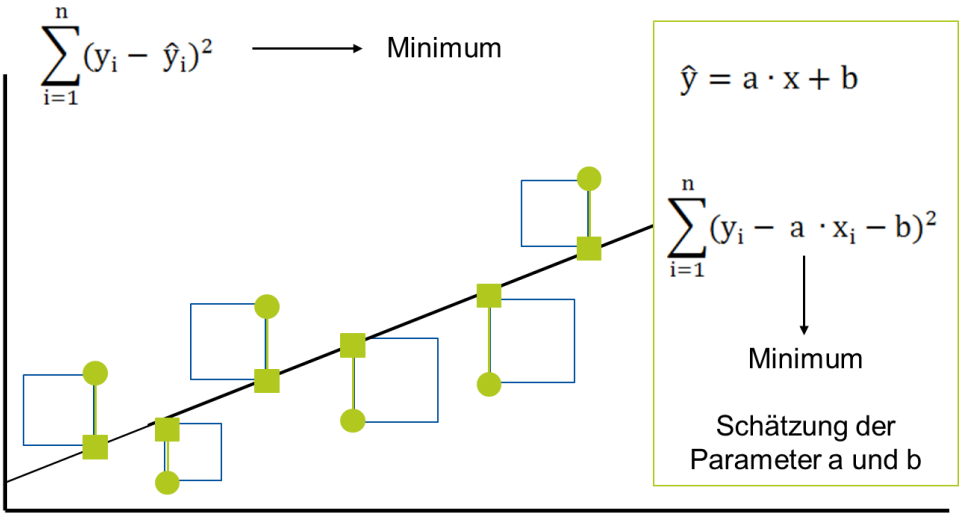


Abbildung 9: Schematische Darstellung der Methode der kleinsten Abweichungsquadrate zur Ermittlung einer linearen Regressfunktion

Dabei können unterschiedlichste Basisfunktionen für die Korrelationsfunktion herangezogen werden (Abbildung 10).

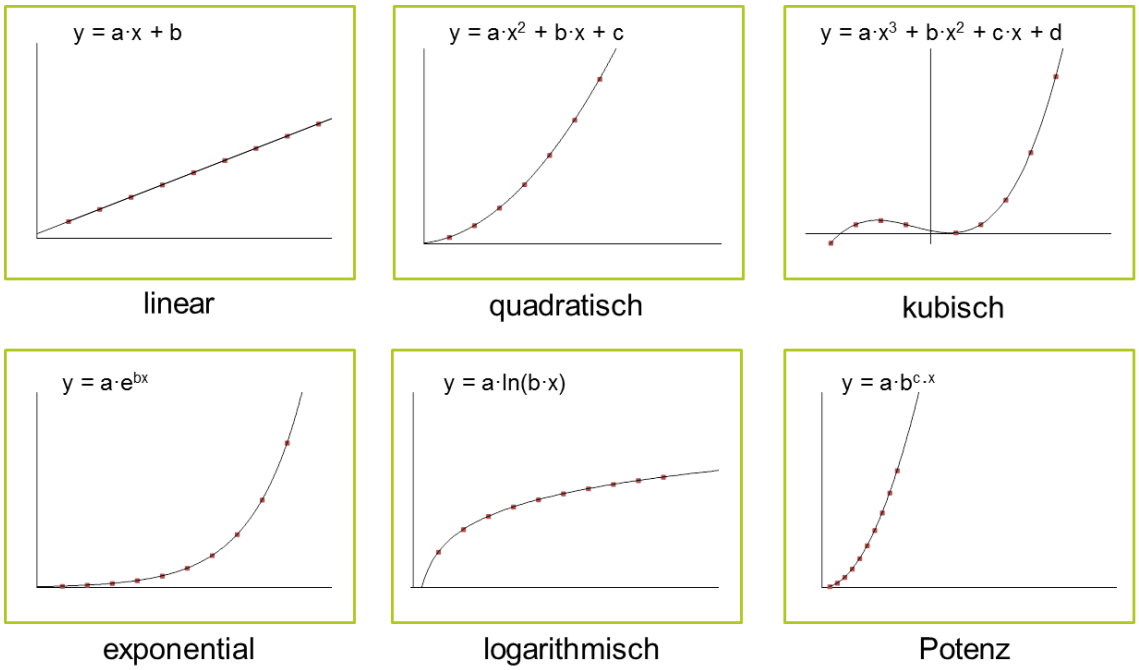


Abbildung 10: Auswahl an möglichen Basisfunktionen für Korrelationsfunktionen

Ist der grundsätzliche funktionale Zusammenhang erkannt, kann mithilfe von Microsoft-Excel meist automatisch die genauste Korrelationsfunktion gefunden werden (Abbildung 11).

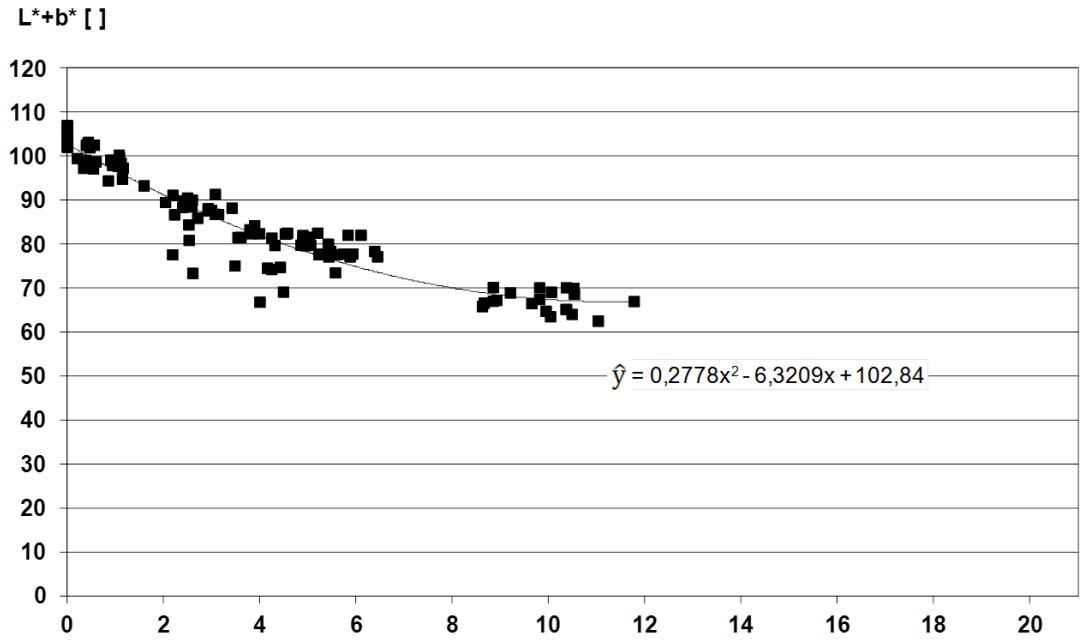


Abbildung 11: Korrelation unter Angabe einer Funktion durch eine Punktwolke aus Messpaaren zweier Merkmale (L*+ b* und dm), eingezeichnet in Koordinatensystem

1.3.2.2 Bestimmtheitsmaß

Die Anpassungsgüte einer mittels der Methode der kleinsten Quadrate ermittelten Ausgleichsfunktion an einen gegebenen Datensatz wird meist über das Bestimmtheitsmaß bewertet.

Gleichung 8: Berechnung des Bestimmtheitsmaßes

Bestimmtheitsmaß

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(y_i - \overline{Y})^2} = \frac{s_{\hat{y}}^2}{s_y^2} \qquad 0 \leq R^2 \geq 1$$

$\hat{y}_i$; durch Regressionsgleichung bestimmte y-Werte

$\overline{Y}$; Arithmetisches Mittel der gemessenen y-Werte

$s_{\hat{y}}^2$; Standardabweichung der durch Regressionsgleichung bestimmten y-Werte

s_y^2; Standardabweichung der gemessenen y-Werte

Das Bestimmtheitsmaß nimmt aufgrund seiner Definition durch den Quotienten der beiden Standardabweichungen lediglich Werte zwischen 0 und 1 an. Ist ei-

ne hohe Anpassungsgüte der Regressionsfunktion an die gemessenen Werte gegeben, so geht der Wert des Bestimmtheitsmaßes gegen 1. ist jedoch keine korrelation zwischen den Wertpaaren erkennbar, so wird das Bestimmtheitsmaß gegen 0 gehen (Abbildung 12).

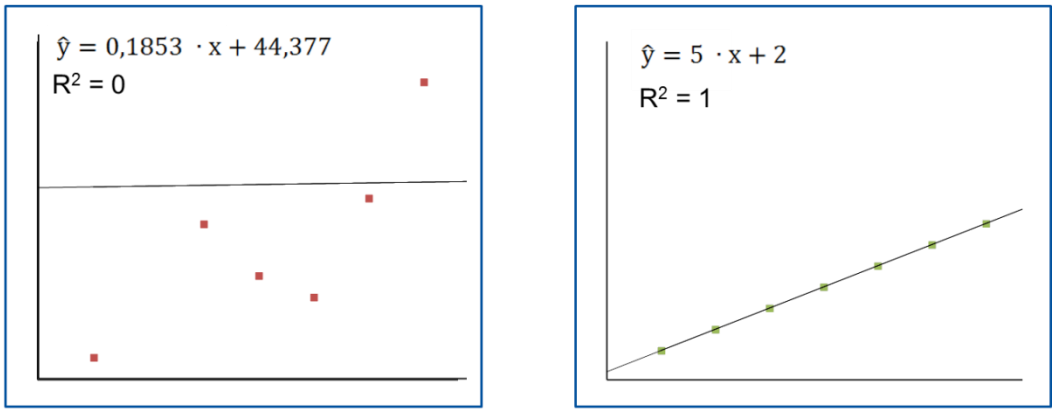


Abbildung 12: Beispielhafte Darstellung unterschiedlicher Bestimmheitsmaße und Korrelationsfunktionen

Werden in der Bachelor- oder Masterarbeit Korrelationen vorgenommen, so sollte stets das Bestimmtheitsmaß mit angegeben werden (Abbildung 13).

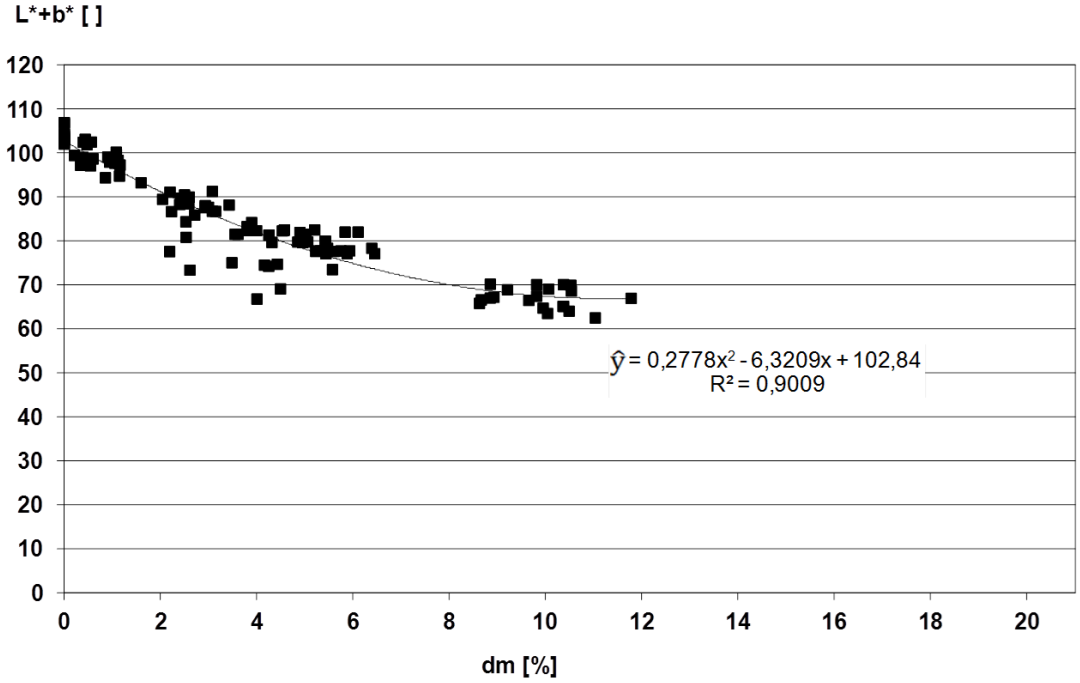


Abbildung 13: Angabe des Bestimmtheitsmaßes einer Korrelation unter Angabe einer Funktion durch eine Punktwolke aus Messpaaren zweier Merkmale (L*+ b* und dm), eingezeichnet in Koordinatensystem