10 Statistik

Diese Kapitel ist als wiederholende Ergänzung vorgesehen. Die wesentlichen Aspekte wurden bereits im Fach Messtechnik und Physikpraktikum vermittelt. Neu ist das Bestimmen der linearen Regressionsfunktion aufgrund der statistischen Masszahlen Mittelwert $\bar{x}$ und Standardabweichung σ.

10.1 Begriffe

Statistik beschäftigt sich mit dem wissenschaftlichen Zusammenfassen, Sammeln, Auswerten und Darstellen von Daten, sowie dem Ziehen von Schlüssen und Treffen von vernünftigen Entscheiden aufgrund solcher Analysen.

Die statistischen Methoden stellen also das Werkzeug zur Behandlung der Daten dar. Im umgangssprachlichen Sinn versteht man unter einer Statistik oftmals eine Auswertung von Daten oder abgeleiteter Grössen. Hier spricht man dann richtigerweise von einer *Unfallstatistik*, *Arbeitslosenstatistik*, etc.

Wir beschäftigen uns mit einem Teilgebiet der Statistik, nämlich der angewandten Statistik: Hier wird konkret mit statistischen Methoden Material verarbeitet im Gegensatz zur theoretischen Statistik. Der von uns betrachtete Teil lässt sich wiederum in zwei Teilgebiete aufteilen:

Beschreibende Statistik (deskriptive Statistik):
Hier werden Daten zusammengefasst und präsentiert. Damit versucht man einen Überblick über das betrachtete Phänomen zu gewinnen oder den Datensatz mit anderen Gruppen zu vergleichen.

Schliessende Statistik (deduktive Statistik):
Da in der beschreibenden Statistik fast ausschliesslich Stichproben bearbeitet werden, versucht man mit geeigneten Methoden Rückschlüsse auf die Grundgesamtheit zu ziehen. Dazu müssen die Stichproben repräsentativ für die Grundgesamtheit sein. Die Beurteilung der Repräsentativität ist eine der schwierigsten Aufgaben der angewandten Statistik.

Wir setzen die Schwerpunkte vor allem im Vermitteln der statistischen Grundbegriffe und Präsentation der Daten.

10.2 Das Histogramm

Das Histogramm ist eine grafische Darstellung der Daten als Balkendiagramm in einer geeigneten Auflösung. Zur Erstellung eines Histogrammes werden mindestens ca. 30 Datenwerte benötigt. Nachfolgend wird beschrieben wie konkret solche Histogramme erstellt werden. Das Vorgehen ist allgemein so üblich und kann als Rezept verstanden werden:

- Man bildet Intervalle oder Klassen, die den Datenwerten xi angepasst sind und bestimmt die Anzahl Werte, die in den Klassen liegen. Hilfsmittel: Strichliste (Häufigkeitstabelle).
- Die Klassenbreite ist (meist) konstant.
- Grundsätzlich gilt: Die Klassenbreiten sind so zu wählen, dass keine Werte auf Klassengrenzen fallen.
  Ist dies nicht nur mit unvertretbarem Aufwand möglich gilt die Methode: Ist ein Wert genau auf einer Klassengrenze, so wird je die Hälfte zur unteren Klasse und oberen Klasse gezählt.
- Die Anzahl der Klassen wird zwischen 5 und 20 gewählt, je nach Umfang der Daten.
- Die Klassenmitten sollten einfache Zahlen sein.
- Die Anzahl *x*-Werte in jedem Intervall trägt man als absolute Häufigkeit oder relative Häufigkeit (in %) auf der Ordinate (y-Achse) auf.

Wir betrachten das Vorgehen zum Erstellen eines Histogrammes anhand eines konkreten Beispiels aus [1] S.14, Brenndauer von Glühlampen:

Um Aufschluss über die Brenndauer von Glühlampen zu erhalten, hat man eine Stichprobe im Umfang von n=90 gezogen und die Brenndauer gemessen. Die 90 Ergebnisse sind in der Urliste (Wertetabelle) festgehalten worden, die hier nicht wiedergegeben sind. Als kürzeste Brenndauer hat man 517h und als längste 1571h ermittelt. Dieses Zahlenmaterial soll nun bearbeitet werden.

10.2.1 Klasseneinteilung

Wir nehmen nun die Klasseneinteilung vor: Wir bestimmen die kürzeste und die längste Brenndauer (hier 517h und 1571h) und wählen eine Klassenbreite so, dass wir etwa 10 bis 20 Klassen erhalten. Im vorliegenden Fall legen wir die Klassenbreite auf 100h fest. Die Klassengrenzen werden auf ganze Hunderte festgelegt (andere Festlegung auch denkbar), so erhalten wir die erste Klasse von 500h..600h, etc.
In einer Strichliste stellen wir zusammen wie viele Elemente in jede Klasse fallen. Aus der Urliste haben wir erhalten:

Brenndauer	Absolute Häufigkeit		Relative. Häufigkeit
500h..600h	///	3	3.3%
600h..700h	////	4	4.4%
700h..800h	///// /	6	6.7%
800h..900h	///// ///// ///	13	14.4%
900h..1000h	///// ///// ///// /////	20	22.2%
1000h..1100h	///// ///// ///// ////	19	21.1%
1100h..1200h	///// ///// ///	13	14.4%
1200h..1300h	///// ///	8	8.9%
1300h..1400h	/	1	1.1%
1400h..1500h	//	2	2.2%
1500h..1600h	/	1	1.1%

10.2.2 Relative Häufigkeit

Wird die zur entsprechenden Klasse gehörende absolute Häufigkeit durch den Umfang der Stichprobe n dividiert, erhalten wir die relative Häufigkeit. Diese wird normalerweise in % oder ‰ angegeben.

Liegt nun die Klasseneinteilung vor, so können wir jede Klasse durch ihre Klassenmitte charakterisieren. Die Klassenmitte der ersten Klasse ist somit 550. Üblicherweise werden die Klassenmitten von *k*-Klassen mit $x_1,..x_k$, oder allgemein mit x_i (i=1..k).

Drücken wir nun die Merkmale der Elemente der Stichprobe durch die entsprechenden Klassenmitten x_i aus, so können wir mit Hilfe der oben gegebenen Berechnungsvorschrift für die relative Häufigkeit eine Funktion *h(x)* definieren , welche jeder Klassenmitte die zugehörige relative Häufigkeit zuordnet. Diese Funktion heisst **Häufigkeitsfunktion:**

$$h(x_i) := \frac{n_i}{n}$$

n_i: Umfang der Klasse i $(i = 1,..,k)$
n: Umfang der Grundgesamtheit
k: Anzahl Klassen

Relative Häufigkeitsfunktion (10-1)

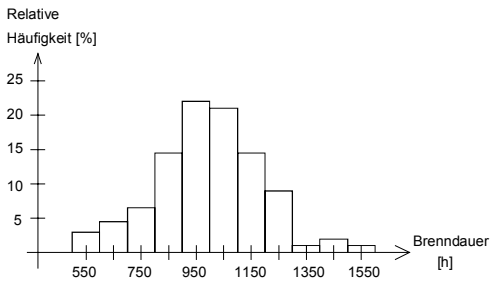
Die Häufigkeiten werden oft in grafisch, im sog. Histogramm, dargestellt.

10.2.3 Grafische Darstellung

Üblicherweise wird nicht der Graph der Häufigkeitsfunktion gezeichnet. Dies hat verschiedene Gründe: Einer ist, dass die Häufigkeitsfunktion in der vorliegenden Form keine stetige Funktion ist , so dass entweder zuerst eine Ausgleichsfunktion bestimmt werden müsste (aufwendig!). Natürlich könnte der Graph als stückweise Funktion gezeichnet werden, allerdings ist die Aussagekraft nicht besonders gross.

Aussagekräftiger ist das Histogramm. Dabei werden auf der Abszisse die Klassenmitten markiert und anschliessend über die Klassenbreite Rechtecke gezeichnet.

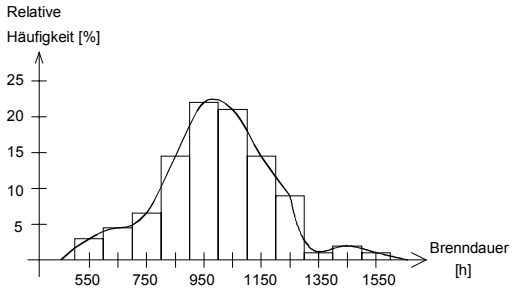
Ist die Klassenbreite konstant, so sind auch die Höhen der Rechtecke proportional zur Häufigkeit. Deshalb werden auf der Ordinate die relative oder absolute Häufigkeit skaliert.



Histogramm mit relativen Häufigkeiten

Je nach der Art der Häufigkeit spricht man von Histogrammen, Prozentsatz-Histogrammen oder relativen Häufigkeitshistogrammen.

Ist der Umfang der Stichprobe genügend gross und wurde eine grosse Anzahl Klassen gewählt, so kann im Histogramm der Treppenzug (Häufigkeitspolygon) durch eine glatte Kurve angenähert werden.



Häufigkeitspolygon im Histogramm durch eine glatte Kurve angenähert.

10.2.4 Relative Summenhäufigkeit, Summenkurve

Stellt man die Frage wie viele Glühlampen eine Brenndauer von höchstens 1000h haben, muss man die absoluten Häufigkeiten für alle Klassen bis 1000h aufsummieren. Dies ergibt 3+4+6+13+20=46. Diese Summe heisst die absolute Summenhäufigkeit.

Die relative Summenhäufigkeit wird Summation der relativen Häufigkeiten errechnet: 3.3%+4.4%+6.7%+14.4%+22.2%=51%. Damit die Rundungsfehler nicht kumulieren, wird die relative Summenhäufigkeit über die absolute Summenhäufigkeit bestimmt.

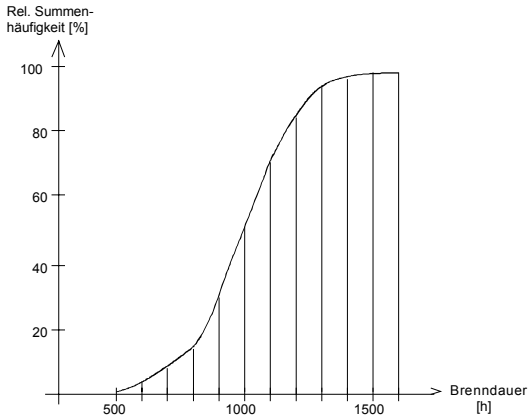
Brenndauer [h]	Klassenmitte x_i	Abs. Häufigkeit n_i	Rel. Häufigkeit n_i/n	Abs. Summen-häufigkeit	Rel. Summen-häufigkeit
500h..600h	550	3	3.3%	3	3.3%
600h..700h	650	4	4.4%	7	7.8%
700h..800h	750	6	6.7%	13	14.4%
800h..900h	850	13	14.4%	26	28.9%
900h..1000h	950	20	22.2%	46	51.1%
1000h..1100h	1050	19	21.1%	65	72.2%
1100h..1200h	1150	13	14.4%	78	86.7%
1200h..1300h	1250	8	8.9%	86	95.6%
1300h..1400h	1350	1	1.1%	87	96.7%
1400h..1500h	1450	2	2.2%	89	98.9%
1500h..1600h	1550	1	1.1%	90	100.0%

Die Summenhäufigkeiten werden den Klassengrenzen zugeordnet. Dies ist klar, da sich die Häufigkeit einer Klasse auf die gesamte Klasse bezieht und somit eine Aussage bis zur Klassengrenze macht. Tabellarisch kann die *Summenhäufigkeitsfunktion* unseres Beispiels beschrieben werden:

500	600	700	800	900	1000'	1100	1200	1300	1400	1500	1600
0%	3.3%	7.8%	14.4%	28.9%	51.1%	72.2%	86.7%	95.5%	96.7%	98.9%	100.0%

Aus der Tabelle können wir nun herauslesen wie viele Glühlampen eine Brenndauer von 1100h haben: 72.2%. So ist dies als 'relative Summenhäufigkeit bis ..' zu verstehen. Dieser 'relativen Summenhäufigkeit bis' sagt der Statistiker *p-Quantil* (oder p%-Quantil).

Oft wird diese Tabelle auch als Graph (geglättet) dargestellt. Man spricht dann von der Summenkurve. Dazu trägt man die Klassengrenzen auf der Abszisse und die relativen Häufigkeiten als Ordinatenwerte auf. Die einzelnen Punkte werden durch Geraden (Summenhäufigkeitspolygon) oder eine glatte Kurve verbunden.



Summenhäufigkeitspolygon durch eine Summenhäufigkeitskurve angenähert

10.3 Mittelwert und Standardabweichung klassierter Daten

Sind die Datenwerte in Klassen erfasst worden, so können Mittelwert und Standardabweichung als *empirischer Grössen* mit den Klassenmitten bestimmt werden.

10.3.1 Mittelwert

Der Mittelwert ist der Lageparameter. Er verkörpert für unser Beispiel die durchschnittliche Brenndauer der Glühlampen.
Dazu die Definition des empirischen Mittelwerte für klassierte Daten:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{k} n_i x_i$$

x_i: Klassenmitte i
n_i: # Merkmale in der Klasse i
k: # Klassen

Empirischer für klassierte Mittelwert (10-2)

Für unser Beispiel ergibt dies konkret: $\bar{x} = \frac{1}{90}(3 \cdot 550 + 4 \cdot 650 + \cdots + 1 \cdot 1550)h = 994h$

Liegt keine Klasseneinteilung vor wird der Mittelwert normal als arithmetisches Mittel berechnet.

10.3.2 Standardabweichung

Die *empirische Standardabweichung klassierter Daten* definieren wir aufgrund der Merkmalswerte in den Klassen:

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

n: # Merkmale der Stichprobe
k: # Klassen
n_i: # Merkmale der Klasse i
x_i: Klassenmitte der Klasse i

Empirische Standardabweichung klassierter Daten (10-3)

Die Division durch *(n-1)* hat theoretische Gründe[1]. Die Standardabweichung ist die Quadratwurzel der Varianz.

[1] Ohne näher auf die Problematik einzugehen sei erklärt:
Früher hat man meist die Standardabweichung definiert als $s := \sqrt{\frac{1}{n}\sum_i n_i (x_i - \bar{x})^2}$. Man erhält dadurch im Mittel etwas zu kleine Werte für die Standardabweichung, verglichen mit Standardabweichung der Grundgesamtheit. Deshalb wird bei Stichprobenrechnungen (n-1) dividiert
Wird hingegen die Standardabweichnung der Grundgesamtheit bestimmt, muss mit n dividiert werden.

Beispiel:
Wir bestimmen nun s für unser Beispiel mit den Glühlampen:

$$s^2 = \frac{1}{89}\left[3(550-994)^2 + 4(650-994)^2 + \cdots + 2(1450-994)^2 + (1550-994)^2\right]h^2 = 4312.5$$
$$\Rightarrow s = 198h$$

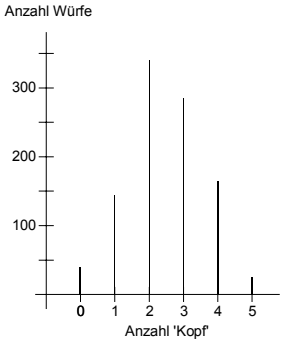
10.4 Stabdiagramme

Stabdiagramme werden vor allem zur Darstellung von Häufigkeitsverteilungen für diskrete Merkmalswerte (-daten) benutzt.

Die Abszisse wird äquidistant skaliert und die Häufigkeiten werden an der Ordinate aufgetragen.

Beispiel:
Fünf Münzen wurden 100mal geworfen und bei jedem wurde die Anzahl 'Kopf' gezählt. Die in der Tabelle festgehaltenen Häufigkeiten werden nun grafisch in einem Stabdiagramm dargestellt.

Anzahl Kopf	Anzahl Würfe (Häufigkeit)
0	38
1	144
2	342
3	287
4	164
5	25
Summe	**1000**



Diskrete Daten können auch mit Histogrammen dargestellt werden. Man legt die Klassenmitten auf die Merkmalswerte. Ein Stabdiagramm sollte nicht mehr als max. 20 Stäbe aufweisen. Wenn mehr Merkmalswerte erfasst wurden, sollten sie in Klassen in einem Histogramm dargestellt werden.

10.5 Aufgaben

Histogramme

1. Die nachfolgende Tabelle zeigt eine Stichprobe aus einem Los Radioröhren. Unter Bezugnahme auf die Daten in der Tabelle bestimme man:

   - Klassenbreite
   - Summe
   - Häufigkeit der vierten Klasse
   - relative Häufigkeit der fünften Klasse
   - Prozentsatz der Röhren mit einer Lebensdauer < 600h
   - Prozentsatz der Röhren mit einer Lebensdauer von mindestens 600h aber weniger als 1000h
   - Statistische Masszahlen

Lebensdauer [h]	Anzahl Röhren
300-399	14
400-499	46
500-599	58
600-699	76
700-799	68
800-899	62
900-999	48
1000-1099	22
1100-1199	6

2. Man konstruiere mit den Daten aus 1.) das Histogramm und das Summenhäufigkeitspolygon.

3. Die nachfolgende Tabelle zeigt die Häufigkeitsverteilungen der Endnoten (in Punkten) in Mathematik und Physik. Bezugnehmend auf die Tabelle bestimme man:

   a.) Den Prozentsatz der Studenten, die 70-79 Punkte in Mathematik und 80-89 Punkte in Physik erhielten.
   b.) Den Prozentsatz der Studenten mit Mathematiknoten unter 70.
   c.) Die Anzahl der Studenten die 70 oder mehr Punkte in Physik und weniger als 80 Punkte in Mathematik erhielten.
   d.) Den Prozentsatz der Studenten, die sowohl in Mathematik wie auch in Physik bestanden haben, wenn zum Bestehen mindestens je 60 Punkte erreicht werden mussten.

		Mathematiknoten					
		40-49	50-59	60-69	70-79	80-89	90-99
Physiknoten	90-99				2	4	4
	80-89			1	4	6	5
	70-79			5	10	8	1
	60-69	1	4	9	5	2	
	50-59	3	6	6	2		
	40-49	3	5	4			