Nützliche Statistikfunktionen in OpenOffice.org Calc

Bereits bestens bekannt sind: **=MITTELWERT()**, **=MAX()**, und **=MIN()**, für die Beschreibung einer Verteilung. NEU sind:

=MODALWERT() für den Modalwert (auch Modus oder häufigster Wert). Hat man 4 Schüler, die 160 cm, zwei mal 170 cm und ein mal 180 cm groß sind, so zählt man im Beispiel wie oft jede Körperlänge insgesamt auftritt, d.h. 160 cm und 180 cm treten nur einmal auf, 170 cm gibt es zwei mal, also ist 170 cm der häufigste Wert, d.h. Modalwert = 170 cm.

=MEDIAN() für den Medianwert (auch Median oder Zentralwert): man ordnet im Beispiel alle Körperlängen der Länge nach, also von 160 cm bis 180 cm und sucht den zentralen Wert, bzw. den Mittelwert der beiden zentralen Werte in dieser Ordnung. Im letzten Beispiel ist die Ordnung: 160 cm, 170 cm, 170 cm, 180 cm der Länge nach und die Mitte liegt genau zwischen 170 cm und 170 cm. Der Mittelwert hiervon ist natürlich wieder 170 cm, so dass der Median = 170 cm ist.

MITTELWERT, MODALWERT und MEDIAN nennt man Lagemaße einer Verteilung, MAX ist das Maximum (der größte Wert) und MIN ist das Minimum (der kleinste Wert) der Verteilung.

=anzahl2() zählt die Anzahl der Argumente.Dabei werden selbst Texteinträge berücksichtigt, die eine leere Zeichenfolge der Länge 0 enthalten. =ANZAHL2(2;4;6;"acht") = 4

Dagegen werden bei ANZAHL(Wert1; Wert2; ... Wert30) nur Zahlen gezählt, Texteinträge bei der Bestimmung der Anzahl nicht berücksichtigt. =ANZAHL(2;4;6;"acht") = 3. Die Anzahl von Zahlen ist folglich 3. Da häufig Hilfsspalten mit laufenden Nummern verwendet werden, ist diese Calc-Funktion oft entbehrlich. Für die Darstellung von Verteilungen muss man Klassen bilden, in dem man eine HÄUFIGKEITsverteilung bildet bzw. **zählt, wenn** bestimmte Kriterien erfüllt sind:

=ZÄHLENWENN() gibt die Anzahl der Zellen in einem Zellbereich zurück, die ein bestimmtes Kriterium erfüllen. A1:D1 sei ein Zellbereich, der die Zahlen 160,170,180,170 enthält. =ZÄHLENWENN(A1:D1;160) ergibt 1. Bei „Texten“ mit Anführungsstrichen arbeiten!

=ZÄHLENWENN(A1:A10;">=170") ergibt 3. Mit dieser Funktion kann man also Tabellen mit Häufigkeiten erstellen. Für die Erstellung der Häufigkeitsverteilung eigentlich vorgesehen ist die Funktion:

=HÄUFIGKEIT(Daten; Klassen) gibt die Häufigkeitsverteilung in einem einspaltigen Array an. Die Funktion zählt die Anzahl der Werte im Array "Daten", die innerhalb der vom Array "Klassen" vorgegebenen Werte liegen. **Daten** stellt die Referenz auf die zu zählenden Werte dar. **Klassen** stellt das Array der Grenzwerte dar. Dies ist eine MATRIXfunktion, siehe Beispiel:

160	170	180	170
Klasseneinteilung:			
<=160 cm	160	1	
> 160 cm bis 175 cm	175	2	
>175 cm		1	

Die 1. Zeile „160...“ enthält die Rohdaten. Die Klassengrenzen sind auf 160 und 175 gesetzt. In C4:C6 steht die Matrixfunktion **{=HÄUFIGKEIT(A1:D1;B4:B6)}** und liefert die absoluten Häufigkeiten 1, 2 und 1 der entsprechenden Klasseneinteilung. (Tastenkombination: Strg+Umschalt+Eingabetaste)

s.a. http://www.tutoria.de/wiki/mathematik/1274/statistische-lagemasse und das **Lehrbuch**.

Unter einer „**Statistik**” versteht man üblicherweise eine Tabelle oder graphische Darstellung von Zahlenmaterial. Daneben bezeichnet das Wort „Statistik” auch die Lehre von den mathematischen Methoden zur Gewinnung und Auswertung von Daten aus statistischen Erhebungen.
Die **beschreibende Statistik** lässt sich innerhalb dieser Lehre abgrenzen, wenn man sich die Arbeitsschritte vergegenwärtigt, die bei der Bearbeitung eines statistischen Problems in der Regel durchlaufen werden.

Diese Schritte sind:

1. Planung der Erhebung

   Vor Beginn der Erhebung muss die Fragestellung präzise gefasst und gegebenenfalls eingeengt werden. Insbesondere ist festzulegen, wie die Daten zu gewinnen sind, mit deren Hilfe man eine Antwort auf die gestellten Fragen geben möchte. Dies geschieht durch Angabe der zu betrachtenden Individuen bzw. Vorgänge, Merkmale und Merkmalsausprägungen.

2. Datenerfassung

   Die Ausprägungen der interessierenden Merkmale (z.B. Temperatur, Länge, Haarfarbe) werden festgestellt und notiert.

3. Datenaufbereitung

   Die erhobenen Daten werden geordnet, in Tabellen oder Graphiken zusammengefasst und dargestellt. Zur kurzen Beschreibung werden Kennzahlen (Mittelwerte, Standardabweichung) berechnet und mögliche Zusammenhänge untersucht (Regressionsrechnung).

4. Auswertung

   Aus den Kenntnissen, die man aufgrund der erhobenen Daten gewonnen hat, werden Schlüsse gezogen und Entscheidungen getroffen.

Die beschreibende Statistik befasst sich vor allem mit der Datenaufbereitung. Im Zusammenhang damit ist unter anderem von Bedeutung, in welcher Form das Datenmaterial dargestellt wird, welche Kennzahlen angegeben werden und wie diese interpretiert werden. So können durch geeignete Akzentuierungen dem Leser einer Statistik bestimmte Informationen bevorzugt vermittelt oder vorenthalten und sogar Fehlinterpretationen suggeriert werden.

Als Werkzeuge verwenden wir das Freewareprogramm GrafStat und die Tabellenkalkulation OpenOffice.org Calc. Wichtige Begriffe der beschreibenden Statistik sind:

· Die Gesamtheit der Individuen oder Objekte, die Gegenstand einer statistischen Untersuchung sind, wird als **Grundgesamtheit** bezeichnet.

· **Stichproben** sind Teilmengen der Grundgesamtheit.

· Die in einer Stichprobe erfassten Daten nennt man allgemein auch **Beobachtungswerte** und bezeichnet sie mit x_i $\{i = 1, 2,\ldots, n\}$.

· Die Eigenschaften der Elemente der Stichprobe werden als **Merkmale** bezeichnet.

· Ein Merkmal kann in mehreren **Merkmalsausprägungen** vorkommen:

- Merkmale mit **Nominalskala** haben als Merkmalsausprägung Namen oder Bezeichnungen, die lediglich der Kennzeichnung dienen (Konfession, Beruf, Haarfarbe)
- Eine Rangskala oder **Ordinalskala** ist Kennzeichen von Merkmalen, deren Ausprägungen eine eindeutige Rangfolge haben (Schulnoten, militärische Rangbezeichnung)
- Merkmale mit **metrischer Skala** haben Merkmalsausprägungen, die sich als

diskrete Zahlenwerte (abzählen) oder auch als stetige Zahlenintervalle (messen) auftreten können.

- Während man Merkmale mit Nominal- und Ordinalskala insgesamt als **qualitative** Merkmale bezeichnet, spricht man bei Merkmalen mit metrischer Skala von **quantitativen** Merkmalen.

Wir können die Merkmalsausprägungen in gewissen Klassen einteilen und dann zählen, wie oft die jeweilige Klasse in der Stichprobe erreicht wurde. Diese Zählung führt zur absoluten Häufigkeit der jeweiligen Klasse. Wenn wir die absolute Häufigkeit durch die Anzahl der Stichprobe teilen ergibt sich die relative Häufigkeit.

Hierzu ein **Beispiel**. Theoretisch ist das Merkmal „Größe der Schüler“ in einer Schulklasse metrisch, stetig und reellwertig verteilt. Anton ist z.B. zu einem bestimmten Zeitpunkt 1,700123 m groß (im Rahmen einer hypothetischen Messgenauigkeit) und Berta zur gleichen Zeit 1,69982387 m groß. Natürlich werden wir beide Messdaten in die Klasse „170 cm“ werfen, ja vermutlich werden wir die Klasseneinteilung sogar so vornehmen, dass alle Größen zwischen 1,65 m und kleiner 1,75 m in diese „Klasse 170 cm“ wandern. Die **absolute Häufigkeit** für diese Klasse ist demnach 2, denn wir haben 2 Schüler gezählt, die etwa 170 cm groß sind. Wenn wir jetzt nur noch 2 weitere Schüler, nämlich Charlotte mit 1,5940201 m und Dirk mit 1,8133419 m haben, die wir gemäß unserer Klasseneinteilung in die Klassen „160 cm“ und „180 cm“ verteilen, so ergibt sich folgende Tabelle:

Klasse:	160 cm	170 cm	180 cm
Beinhaltene Daten (Beobachtungswerte)	x_1=1,5940201	x_2=1,700123 und x_3=1,69982387	x_4=1,8133419
Absolute Häufigkeit:	1 (ein Wert bzw. Schüler)	2 (zwei Werte bzw. Schüler)	1 (ein Wert bzw. Schüler)
Relative Häufigkeit:	$\frac{1}{4}=0{,}25=25\ \%$	$\frac{2}{4}=\frac{1}{2}=0{,}5=50\ \%$	$\frac{1}{4}=0{,}25=25\ \%$

Zur Berechnung der **relativen Häufigkeit** wurde die Anzahl der für die jeweilige eingeteilten Klasse günstigen Beobachtungswerte durch die Gesamtzahl aller Beobachtungswerte der Stichprobe (n=4) geteilt.

Die Summe aller relativen Häufigkeiten ergibt 1=100 %.

Gern wird so eine Größenverteilung in Form eines Säulen-, Balken-, Kuchendiagramms oder als Histogramm (Säulendiagramm mit Abstand 0) dargestellt mittels GrafStat oder OpenOffice.org Calc. Die Klasseneinteilung auf der ersten Seite wo explizit die Funktionsaufrufe in Calc dargestellt wurden unterscheidet sich übrigens von der hier dargestellten Klasseneinteilung. Bei den hier verwendeten Beobachtungswerten könnte man drakonisch scharf mit einer Klasseneinteilung gemäß den Klassengrenzen 160 cm und 170 cm folgende (andere!) Verteilung darstellen:

Klasse:	<=160 cm	> 160 cm bis <=170 cm	>170 cm
Beinhaltene Daten (Beobachtungswerte)	1,5940201	1,6998239	1,7001230 und 1,8133419
Relative Häufigkeit:	$\frac{1}{4}=0{,}25=25\ \%$	$\frac{1}{4}=0{,}25=25\ \%$	$\frac{2}{4}=\frac{1}{2}=0{,}5=50\ \%$

Bei Verwendung der Klassengrenzen 160 und 175 haben wir wieder das Ergebnis auf Seite 1 (vorherige Verteilung).