Weiteres zu Statistikfunktionen in OpenOffice.org Calc

Um die Ergebnisse zu präsentieren, bieten sich Grafiken an, wie sie z.B. auf S. 29 im Lehrbuch dargestellt sind. Statistische Kennzahlen sind neben den graphischen Darstellungsmöglichkeiten eine weitere Form der Beschreibung der erhobenen Daten. Eine Tabelle oder Graphik gibt uns Auskunft über die gesamte Verteilung eines Merkmals. Oftmals ist es aber auch sinnvoll, sparsam und mitunter auch übersichtlicher, die Verteilung des Merkmals nicht graphisch darzustellen. Dann benötigen wir andere Formen der Beschreibung der Verteilung: Die statistischen Kennzahlen. Hierbei unterscheiden wir in der deskriptiven Statistik zwischen den folgenden beiden Gruppen.

Lagemaße (Maße der zentralen Tendenz)

Lagemaße sind solche Kennwerte, die alle Messwerte, damit also die Verteilung eines bestimmten Merkmals am besten repräsentieren. Jedes der drei nun vorgestellten Lagemaße hat auf seine Art und Weise die Eigenschaft, die Verteilung „am besten“ zu repräsentieren.

Das **arithmetische Mittel** (Mittelwert, Durchschnitt) ist das wohl bekannteste und am häufigsten verwendete Lagemaß, in CALC schreiben wir =MITTELWERT(), und berechnet sich formal:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad .$$

Arithmetisches Mittel = Summe der Beobachtungen dividiert durch die Anzahl der Beobachtungen.

Es gilt:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Das zugrunde liegende Konzept des arithmetischen Mittels ist es, einen Punkt der Verteilung zu finden, an dem sie sich „im Gleichgewicht“ befindet. Man kann sich die einzelnen Messwerte entlang eines (fiktiven) Stabes aufgetragen vorstellen und versuchen, jenen Punkt zu finden, an dem dieser ausbalanciert ist. Dieser Mittelwert ist relativ empfindlich gegenüber Ausreißern.

Der **Median** oder **Zentralwert** gibt jenen Punkt an, der genau zwischen der oberen und unteren Hälfte der Verteilung liegt, d.h. der Median befindet sich in einer der Größe nach sortierten Liste von Messwerten genau in der Mitte, in CALC schreiben wir =MEDIAN(). Daher hat er die Eigenschaft, dass jeweils mindestens die Hälfte der beobachteten Messwerte einen Wert größer oder gleich (bzw. kleiner oder gleich) dem Median annehmen. Symbolisch:

$$\tilde{x}_{0,5} = \begin{cases} x_{(\frac{n+1}{2})} & \text{sofern die Werte der Größe nach sortiert sind und n ungerade bzw.} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}) & \text{n gerade ist (} \quad x_{(\frac{n+2}{2})} = x_{(\frac{n}{2}+1)} \quad \text{).} \end{cases}$$

Der **Modus** oder **Modalwert** (häufigster Wert) bezeichnet schließlich jenen Wert, der am häufigsten in der Liste der erhobenen Messwerte auftritt. In CALC schreiben wir =MODALWERT().

Median und Modus sind relativ unempfindlich gegenüber Ausreißern, d.h. falls wir den kleinsten oder größten Wert streichen müssen, da wir z.B. das Komma falsch gesetzt haben, passiert bei beiden wenig (u.U. gar nichts).

Alle diese Lagemaße sind bereits in http://www.warncke-family.de/dv/calc/statistik1.pdf beschrieben. Nun zur zweiten Gruppe:

Streuungsmaße (Dispersionsmaße)

Die Motivation, neben den Lagemaßen noch andere Maßzahlen zur Charakterisierung einer Verteilung heranzuziehen, wird anhand Abbildung 1 deutlich. Sie zeigt zwei Verteilungen, deren Lagemaß und Anzahl der Messwerte gleich sind, die aber dennoch vollkommen unterschiedlich aussehen und auch unterschiedliche Interpretationen verlangen.

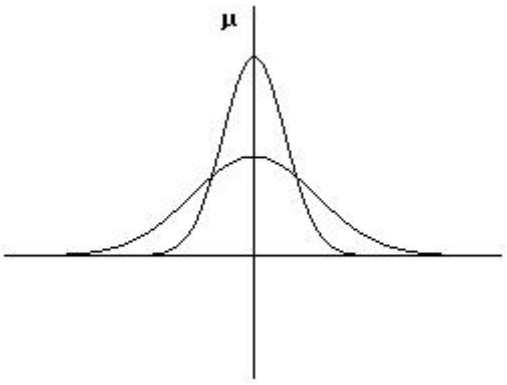

Abb. 1

Man sieht ganz leicht, dass die Werte der „breit auseinander gezogenen“ Verteilung stärker streuen als jene der „nach oben gestreckten“ Verteilung. Nun brauchen wir auch für das Ausmaß der Streuung geeignete statistische Kennzahlen.

Wir unterscheiden vier Streuungsmaße (Dispersionsmaße):

Die **Varianz** ist die mittlere quadratische Abweichung der einzelnen Beobachtungen vom arithmetischen Mittel, sie ist mehr von theoretischer Bedeutung und wird praktisch nicht benutzt, formal:

$$s_{\bar{x}}^{\,2} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \,.$$

Die **Standardabweichung der Grundgesamtheit** ist definiert als die Quadratwurzel aus der Varianz, in CALC schreiben wir =STABWN(), formal:

$$s_{\bar{x}} = \sqrt{s_{\bar{x}}^2} \,.$$

In der Praxis haben wir es nicht mit der theoretischen Grundgesamtheit aller möglichen Daten zu tun, sondern nur mit einer sogenannten Stichprobe (unsere Interviews werden sich z.B. immer nur auf eine kleine Auswahl beschränken). Die **Schätzung der Standardabweichung** erfolgt dann in CALC über =STABW(), formal über die Beziehung

$$\sigma = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Die **Spannweite** ist das einfachste Maß der Streuung und berechnet sich wie folgt:

$R = x_{\max} - x_{\min}$, also Differenz zwischen größtem und kleinstem Messwert, in CALC als =MAX()-MIN(). Es ist natürlich sehr empfindlich gegenüber Ausreißern (vgl. oben). Warum ist dieses Maß nicht besonders aussagekräftig?

Literatur: Lehrbuch von Bohner, K., Ihlenburg, P., Ott, R.: „Beschreibende Statistik“, Rinteln: Merkur Verlag 2003

und http://www.mathe-online.at/clips/mwstdabw/index.html