Stichprobenanalyse

Autor & Copyright: Dipl.-Ing. Harald Nahrstedt

Version: 2016 / 2019 / 2021 / 365

Erstellungsdatum: 28.04.2010

Überarbeitung: 01.12.2023

Quelle: Vorlesungsscript

Beschreibung:

Die Technische Statistik beschäftigt sich unter anderem auch mit Daten, die als Stichproben aus einer Grundgesamtheit bestimmt wurden. Deren Analyse erlaubt dann wiederum Rückschlüsse auf die Merkmale der Grundgesamtheit. Die Stichproben bilden eine Teilmenge der Grundgesamtheit und werden meist nach dem Zufallsprinzip ausgewählt.

Anwendungs-Dateien:

06-15-01_Stichprobenanalyse1.xlsm

06-15-01_Stichprobenanalyse2.xlsm

Zur Bestimmung von Stichproben stehen verschiedene Verfahren zur Verfügung. Die richtige Auswahl der Stichproben ist sehr wichtig, da diese repräsentativ für die Grundgesamtheit stehen. Ein Beispiel dafür sind die Hochrechnungen zu den Wahlen. Die Kombinatorik befasst sich mit den grundlegenden Möglichkeiten von Auswahlverfahren. Stellvertretend soll an dieser Stelle das Zufallsauswahlverfahren behandelt werden. Bereits in meinem Buch habe ich es beschrieben.

1 Verteilungsparameter

Zum besseren Verständnis vorliegender Daten wie Messreihen, ist es oft sinnvoll, diese in Verteilungsparametern zusammen zu fassen. Die Parameter sagen etwas über die Lage der Daten aus. Ebenso dienen sie zur Beschreibung ihrer Ausbreitung um einen mittleren Wert. Wir gehen von einer Menge von n Merkmalswerten (Stichproben) aus

$$x_1, x_2, \dots, x_n \tag{1}$$

und definieren den Mittel- oder Durchschnittswert zu

$$\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2}$$

Es gibt noch zwei andere Parameter zur „Lage“ der Daten, der Zentralwert und der Modalwert. Der Zentralwert z von n Merkmalswerten definiert sich als der Zahlenwert in der Mitte, falls er existiert. Um ihn zu bestimmen, müssen die Daten aufsteigend sortiert vorliegen. Danach gelten für eine ungerade Anzahl Daten

$$z = x_{\frac{n+1}{2}} \tag{3}$$

und für eine gerade Anzahl Daten

$$z = \frac{x_{\frac{n}{2}}+x_{\frac{n+2}{2}}}{2}. \tag{4}$$

Der dritte Parameter zur Lage der Daten ist der Modalwert. Er definiert sich als der Wert, der am häufigsten vorkommt. Eine Verteilung kann somit mehr als einen Modalwert besitzen und wird dann als multimodal bezeichnet. Der Maximal- und der Minimalwert sagen ebenso etwas über die Verteilung der Daten aus wie Variationsbreite b, die sich aus der Differenz der beiden ergibt.

$$b = Max\{x_i\} - Min\{x_i\} \tag{5}$$

Ein sehr gebräuchlicher Parameter ist die Varianz. Sie ist definiert zu

$$s^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] \tag{6}$$

oder in Kurzform

$$s^2 = \frac{1}{n-1}[\sum(x_\mathrm{i} - \bar{x})^2]. \tag{7}$$

Die Quadratwurzel aus der Varianz ist die Standardabweichung. Sie ist ein Maß für die Streuung der Daten um den Mittelwert. Ist die Verteilung der Daten annähernd symmetrisch, so liegen Zentral- und Mittelwert nahe beieinander. Ein Maß für die Asymmetrie (Schiefe) einer Verteilung lautet

$$v = \frac{3(\bar{x}-z)}{s}. \tag{8}$$

Die Asymmetrie kann sowohl negativ wie auch positiv sein, wie die nachfolgende Darstellung anschaulich zeigt (Bild 1).

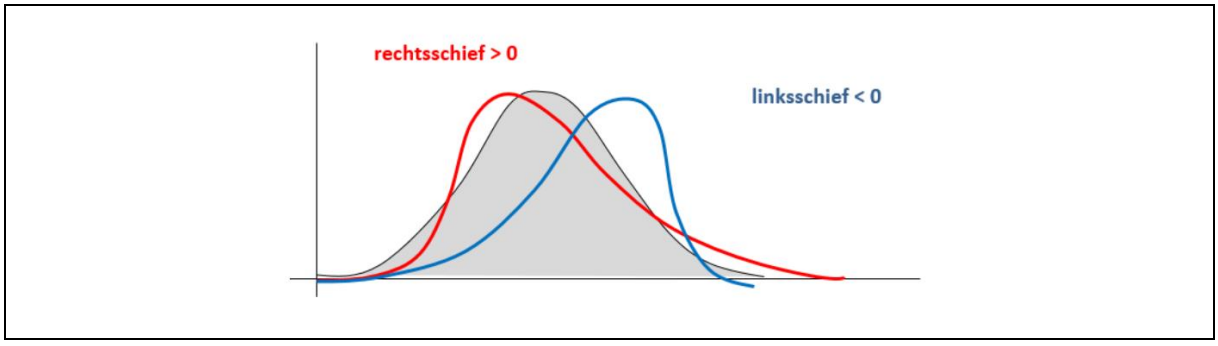

Bild 1. Asymmetrie einer statistischen Verteilung

Als Anwendungsbeispiel wird die Oberflächenrauheit einer Werkstückoberfläche auf einer Messstrecke gemessen, dabei werden die folgenden 50 Profiltiefen in µm ermittelt (Bild 2).

	A	B	C	D	E	F	G	H	I	J
1	138	448	724	418	145	139	316	468	345	333
2	304	131	397	521	313	725	171	168	680	604
3	377	448	388	174	165	278	202	356	112	350
4	388	166	183	225	297	303	404	299	276	341
5	166	513	586	604	534	414	365	310	229	256

Bild 2. Messwerte

Wir gehen zunächst davon aus, dass diese Messwerte in einer Tabelle gespeichert sind, und zwar mit einer beliebigen Anzahl Zeilen und Spalten. Aber immer so, dass sie ein wohlgeordnetes Rechteck bilden. Im Extremfall ist es eine Zeile oder Spalte. Die ermittelten Parameter sollen nach dem Start auf der entsprechenden Datentabelle in einem Formular angezeigt werden.

Für die Programmierung werden ein Modul *modParameter* und ein Formular *frmParameter* eingefügt (Bild 3).

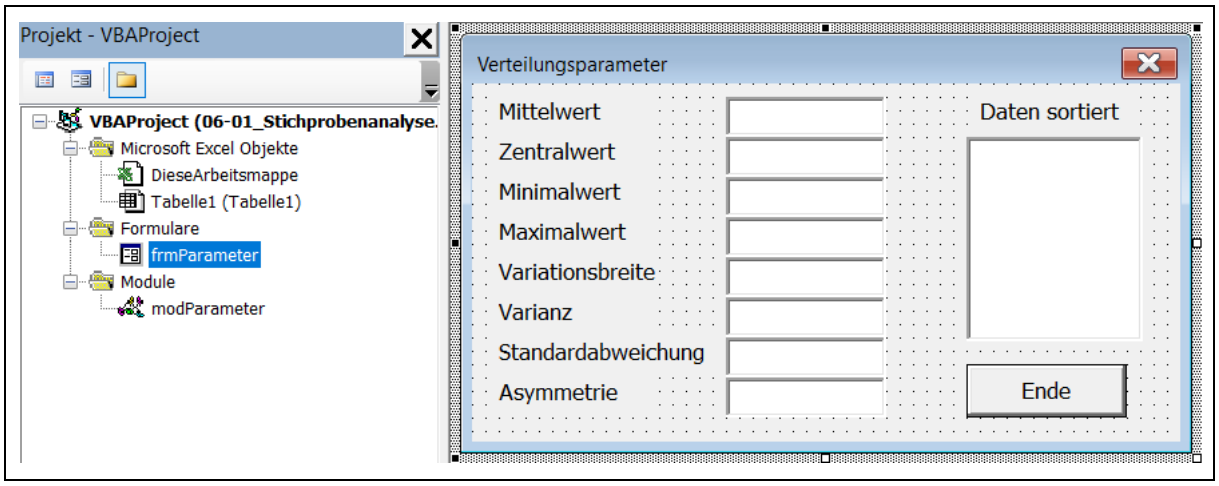

Bild 3. Die Objekte des Projekts und die Ansicht des Formulars

Im Formular wird lediglich die Schaltfläche *cmdEnde* programmiert, um das Ausschalten des Formulars beim Anklicken durchzuführen.

Codeliste 1. Ereignisprozedur im Formular frmParameter

```
Option Explicit

Private Sub cmdEnde_Click()
   Unload Me
End Sub
```

Codeliste 2. Analyseprozedur im Modul modParameter

```
Sub Parameter()
   Dim iMaxRow     As Integer
   Dim iMaxCol     As Integer
   Dim iCR         As Integer
   Dim iCC         As Integer
   Dim iC          As Integer
   Dim iMax        As Integer
   Dim dMit        As Double
   Dim dSum        As Double
   Dim dX          As Double
   Dim dZ          As Double
   Dim dMin        As Double
   Dim dMax        As Double
   Dim dS          As Double
   Dim dss         As Double
   Dim dv          As Double
   Dim Daten()     As Double

'Formular instanziieren
   Load frmParameter
```

```
'Benutzte Zeilen und Spalten
   iMaxRow = ActiveSheet.UsedRange.Rows.Count
   iMaxCol = ActiveSheet.UsedRange.Columns.Count
   iMax = iMaxRow * iMaxCol
   ReDim Daten(iMax) As Double

'Lesen und Summenbildung
   dSum = 0
   dMin = ActiveSheet.Cells(1, 1)
   dMax = ActiveSheet.Cells(1, 1)
   For iCR = 1 To iMaxRow
      For iCC = 1 To iMaxCol
         dSum = dSum + ActiveSheet.Cells(iCR, iCC)
         iC = (iCR - 1) * iMaxCol + iCC
         Daten(iC) = ActiveSheet.Cells(iCR, iCC)
         If ActiveSheet.Cells(iCR, iCC) < dMin Then _
            dMin = ActiveSheet.Cells(iCR, iCC)
         If ActiveSheet.Cells(iCR, iCC) > dMax Then _
            dMax = ActiveSheet.Cells(iCR, iCC)
      Next iCC
   Next iCR

'Mittelwert und Extremwerte
   dMit = dSum / iMax
   frmParameter.tbxMittelwert = Format(dMit, "#0.000")
   frmParameter.tbxMinimalwert = dMin
   frmParameter.tbxMaximalwert = dMax
   frmParameter.tbxVariation = dMax - dMin

'Sortierung
   For iCR = 1 To iMax - 1
      For iCC = iCR + 1 To iMax
         If Daten(iCR) > Daten(iCC) Then
            dX = Daten(iCR)
            Daten(iCR) = Daten(iCC)
            Daten(iCC) = dX
         End If
      Next iCC
   Next iCR
   For iCR = 1 To iMax
      frmParameter.lbxDaten.AddItem Daten(iCR)
   Next iCR

'Zentralwert
   If iMax / 2 = Int(iMax / 2) Then
   'gerade
      dZ = (Daten(iMax / 2) + Daten((iMax + 2) / 2)) / 2
   Else
   'ungerade
      dZ = Daten((iMax + 1) / 2)
   End If
   frmParameter.tbxZentralwert = Format(dZ, "#0.000")

'Varianz
   dSum = 0
   For iCR = 1 To iMax
      dX = Daten(iCR)
      dSum = dSum + (dX - dMit) * (dX - dMit)
   Next iCR
   dss = dSum / (iMax - 1)
   dS = Sqr(dss)
   frmParameter.tbxVarianz = Format(dss, "#0.000")
   frmParameter.tbxStandard = Format(dS, "#0.000")

   'Asymmetrie
   dv = 3 * (dMit - dZ) / dS
   Stichprobenanalysen 5
   frmParameter.tbxAsymmetrie = Format(dv, "#0.000")
'Formular zeigen
   frmParameter.Show
End Sub
```

Die Daten werden ohne Formatierung im Formular ausgegeben (Bild 4).

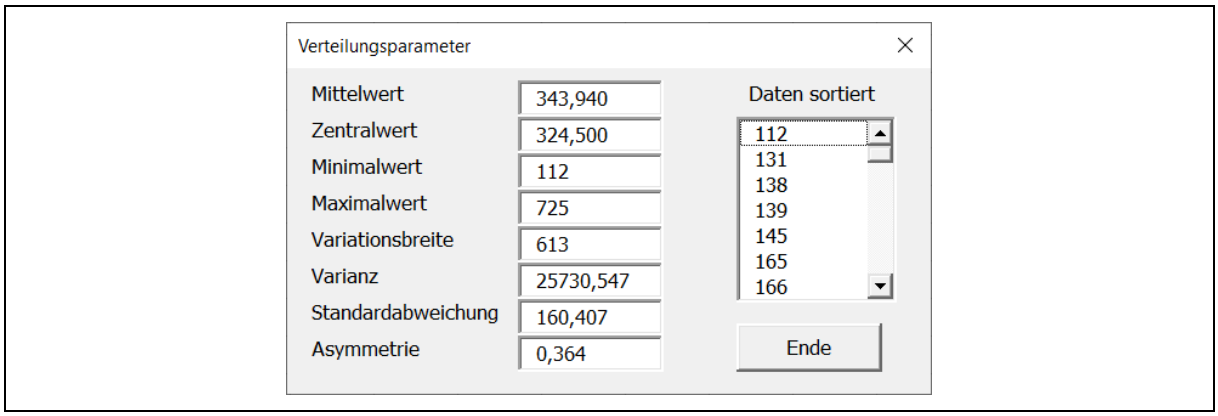

Bild 4. Auswertung der Beispieldaten

2 Regression und Korrelation

Wurde im vorangegangenen Kapitel ein Merkmal einer Datenmenge betrachtet, so sind oftmals mehrere Merkmale gleichzeitig zu analysieren. Dabei müssen Beziehungen zwischen den Merkmalen untereinander gefunden werden. Stichproben der zugehörigen Merkmale können als Punktepaare in einem Diagramm dargestellt werden. Wie die Darstellung zeigt, lassen sich lineare Tendenzen erkennen (Bild 5).

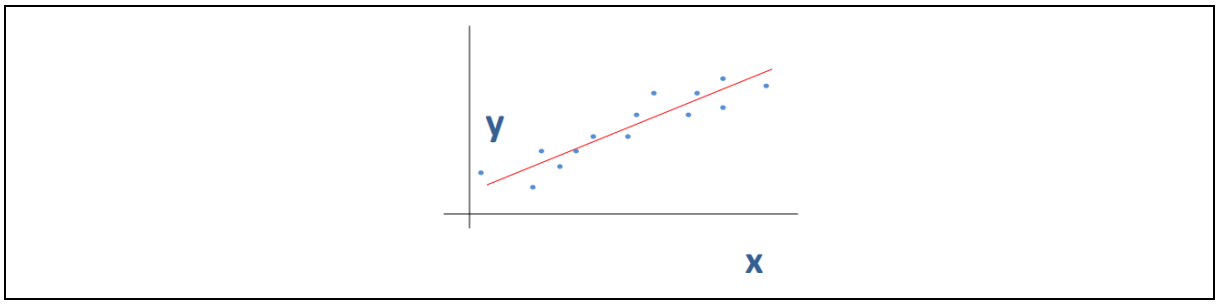

Bild 5. Funktionale Abhängigkeit zweier Merkmale

Doch wie stark ist diese lineare Korrelation zwischen den Merkmalen? Der Grad der linearen Abhängigkeit wird durch den linearen Korrelationskoeffizienten gekennzeichnet, der sich wie folgt definiert:

$$r = \frac{[(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})]}{\sqrt{[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2][(y_1-\bar{y})^2+(y_1-\bar{y})^2+\cdots+(y_1-\bar{y})^2]}}. \quad (9)$$

In Kurzform

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \cdot \sum(y_i-\bar{y})^2}}. \quad (10)$$

Diese Gleichung lässt sich auch umstellen in die Form

$$r = \frac{\frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2 \cdot \frac{1}{n-1}\sum(y_i-\bar{y})^2}}. \quad (11)$$

Somit erhalten wir im Nenner die Quadratwurzel aus dem Produkt der Varianzen von x und y. Den Zähler bezeichnet man als Kovarianz und schreibt verkürzt

$$r = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}}. \quad (12)$$

Per Definition gilt, dass -1 <= r <= 1 ist. Liegt r nahe bei ± 1, so wird damit ein hoher Grad von linearer Abhängigkeit ausgedrückt, während r nahe bei 0 keine lineare Abhängigkeit kennzeichnet. Während r nahe bei 1 eine proportionale Abhängigkeit darstellt, ist r nahe bei -1 eine Aussage über eine umgekehrt proportionale Abhängigkeit. Bei einer Maschine sind Alter und Reparaturkosten sicher direkt proportional, während Alter und Ausfallwahrscheinlichkeit sicher umgekehrt proportional sind.

Liegt eine lineare Abhängigkeit vor, so will man diese durch eine Funktion beschreiben. Die allgemeine Form der Geradengleichung lautet

$$y = a \cdot x + b. \tag{13}$$

Eine sehr häufig angewandte Methode, die optimale Gerade für n Wertepaare (x, y) zu finden, ist die Methode der kleinsten Fehlerquadrate. Bei dieser Methode sucht man die Gerade, bei der der Abstand aller Punkte im Diagramm zur Geraden am kleinsten wird (rote Linien). Und zwar zur Verschärfung wird dieser Wert ins Quadrat gesetzt (grüne Quadrate) (Bild 6).

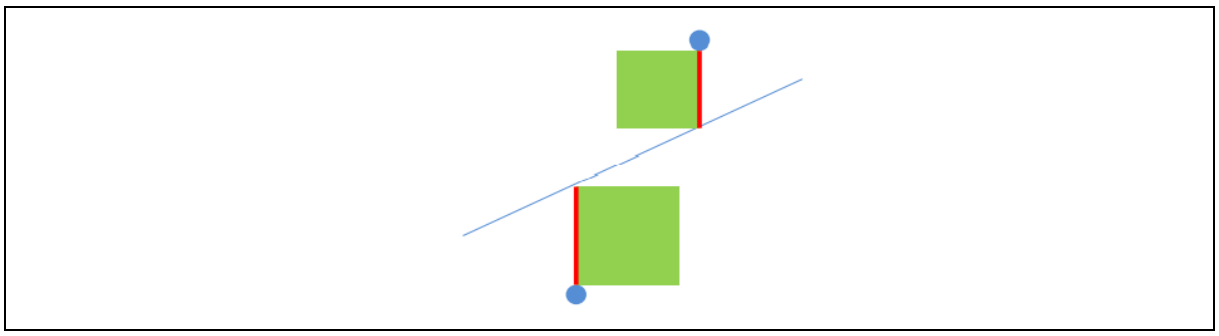

Bild 6. Methode der kleinsten Fehlerquadrate

Mathematisch ausgedrückt erhält man für alle Fehler

$$Min(\sum(y_i - (ax_i + b))). \tag{14}$$

Und natürlich für die Fehler zum Quadrat

$$Min(\sum(y_i - (ax_i + b))^2). \tag{15}$$

Im Idealfall ist

$$\sum(y_i - (ax_i + b)) = 0 \tag{16}$$

und auch

$$\sum(y_i - (ax_i + b))^2 = 0. \tag{17}$$

Umgestellt wird daraus

$$a\sum x_i + n \cdot b = \sum y_i \tag{18}$$

bzw.

$$a\sum x_i^2 + b\sum x_i = \sum x_i y_i. \tag{19}$$

Diese beiden Gleichungen lassen sich analytisch lösen zu

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_1)^2} \tag{20}$$

und

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}. \tag{21}$$

Als weiteres Anwendungsbeispiel betrachten wir die Korrelation zwischen Preis und Nachfrage. Ein Marktforschungsinstitut ermittelt zu einem festgesetzten Zeitraum die folgenden Nachfragewerte eines Produktes. Die Spalte A zeigt den Preis, die Spalte B die Nachfrage (Bild 7).

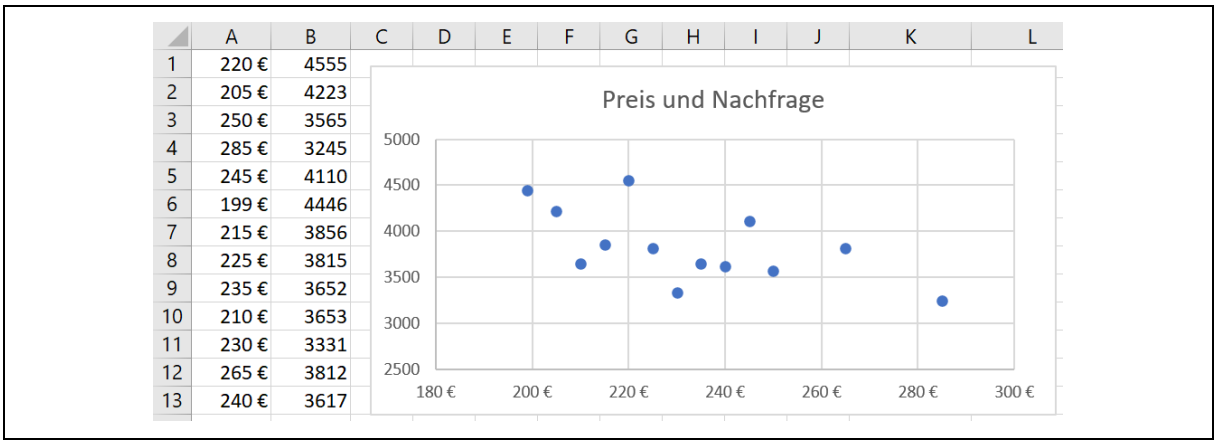

Bild 7. Preis und Nachfrage

Damit das Programm für verschiedene Auswertungen genutzt werden kann, soll die Anzahl der Punktwerte als beliebig vorausgesetzt werden. Ein eingeblendetes Formblatt *frmRegression* (Bild 8) soll wie zuvor die ermittelten Werte dieses Kapitels anzeigen.

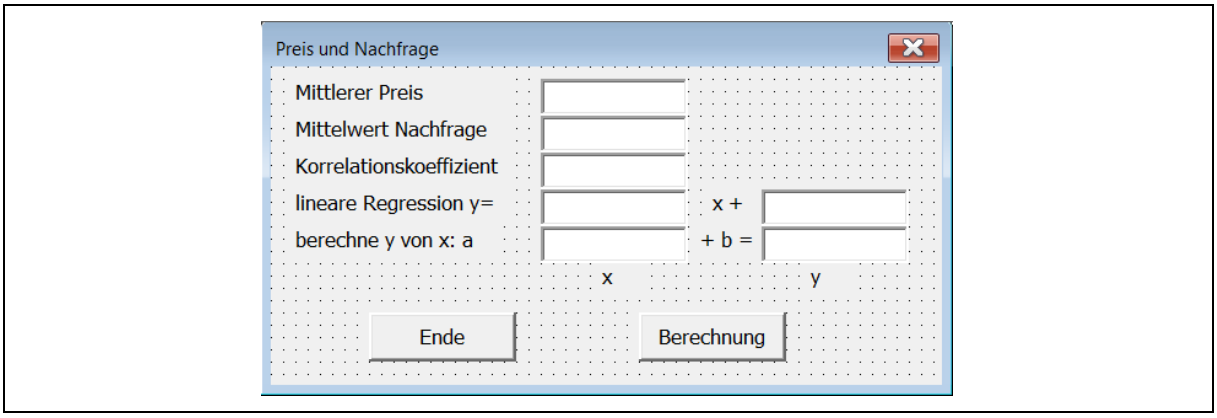

Bild 8. Formular zur Berechnung

Codeliste 3. Prozeduren im Formular frmRegression

```
Option Explicit

'Anzeige schließen
Private Sub cmdEnde_Click()
   Unload Me
End Sub

'y-Wert berechnen
Private Sub cmdBerechnung_Click()
   Dim dA As Double
   Dim dB As Double
   Dim dX As Double
   Dim dY As Double

   dA = tbxA.Value
   dB = tbxB.Value
   dX = tbxX.Value
   dY = dA * dX + dB
   tbxY.Value = Format(dY, "#0.00")
End Sub
```

Codeliste 4. Prozedur im Modul modRegression

```
Option Explicit

Sub Nachfragen()
   Dim n       As Integer
   Dim iCR     As Integer
   Dim dSumX   As Double
   Dim dSumY   As Double
   Dim dSumXX  As Double
   Dim dSumYY  As Double
```

```
    Dim dSumXY  As Double
    Dim dMitX   As Double
    Dim dMitY   As Double
    Dim dA      As Double
    Dim dB      As Double
    Dim x       As Double
    Dim y       As Double
    Dim z1      As Double
    Dim z2      As Double
    Dim z3      As Double
    Dim dKor    As Double

'Formular instanziieren
    Load frmRegression

'Benutzte Zeilen und Spalten
    n = ActiveSheet.UsedRange.Rows.Count

'Lesen und Summenbildungen
    dSumX = 0
    dSumY = 0
    dSumXX = 0
    dSumYY = 0
    dSumXY = 0
    For iCR = 1 To n
        x = ActiveSheet.Cells(iCR, 1)
        y = ActiveSheet.Cells(iCR, 2)
        dSumX = dSumX + x
        dSumY = dSumY + y
        dSumXX = dSumXX + x * x
        dSumYY = dSumYY + y * y
        dSumXY = dSumXY + x * y
    Next iCR

'Mittelwerte
    dMitX = dSumX / n
    frmRegression.tbxMitPrs = Format(dMitX, "#0.00")
    dMitY = dSumY / n
    frmRegression.tbxMitAnz = Format(dMitY, "#0.00")

'Korrelationskoeffizient
    z1 = dSumXY - dSumX * dSumY / n
    z2 = dSumXX - dSumX * dSumX / n
    z3 = dSumYY - dSumY * dSumY / n
    dKor = z1 / Sqr(z2 * z3)
    frmRegression.tbxKorKoe = Format(dKor, "#0.00")

'lineare Regression
    z1 = n * dSumXX - dSumX * dSumX
    dA = (n * dSumXY - dSumX * dSumY) / z1
    dB = (dSumXX * dSumY - dSumX * dSumXY) / z1
    frmRegression.tbxA = Format(dA, "#0.00")
    frmRegression.tbxB = Format(dB, "#0.00")

'Formular zeigen
    frmRegression.Show
End Sub
```

Die Daten werden ohne Formatierung im Formular ausgegeben (Bild 9).

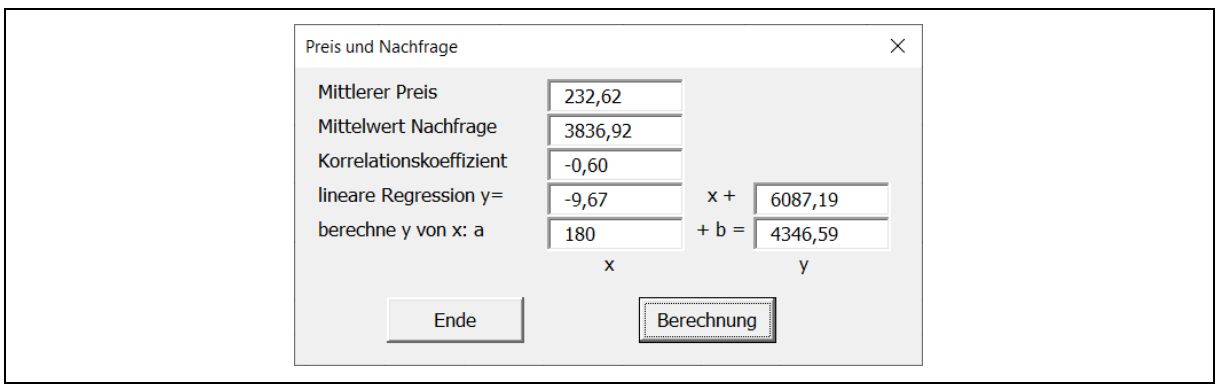

Bild 9. Auswertung der Beispieldaten

Ein weiterer Service erlaubt mittels linearer Regression die Bestimmung möglicher Nachfragewerte in Abhängigkeit vom Preis.

3 Parameterschätzung und Konfidenzintervall

Werden verschiedene Stichproben aus einer Grundgesamtheit analysiert, zum Beispiel in Bezug auf einen Mittelwert, so wird dieser mit jeder Stichprobe anders ausfallen. Die mit einer Stichprobe ermittelten Parameter sind nur Schätzwerte bzw. Näherungswerte der tatsächlichen Parameter der Grundgesamtheit. Man spricht daher von Punktschätzer. Sie sagen jedoch nichts darüber aus, wie sehr sie sich den tatsächlichen Parametern annähern.

Man nähert sich diesem Problem in der Statistik durch die Betrachtung eines Intervalls, dessen Grenzen vorgeben, mit welcher Wahrscheinlichkeit als Prozentsatz, der Punktschätzer in diesem Intervall liegt. Diese Betrachtung heißt Intervallschätzung.

Zur Durchführung der Intervallschätzung gibt es unterschiedliche Methoden, die von der Häufigkeitsverteilung der Stichproben abhängt. Dazu noch ein paar grundlegende Gegebenheiten. Mit der Häufigkeit der Stichproben nähern sich deren Mittelwerte einer Normalverteilung an. Bewiesen durch den Zentralen Grenzwertsatz der Wahrscheinlichkeitstheorie. Ebenso nähern sich die Mittelwerte bei hinreichender Anzahl von Stichproben dem Mittelwert der Grundgesamtheit an. Denn im Grenzfall ist der Mittelwert aller Stichproben mit dem Mittelwert der Grundgesamtheit identisch.

Ebenso ist die Streuung der Mittelwerte der Stichproben geringer als die Streuung der Stichproben selbst, da Ausreißer bei den Stichproben, bei den Mittelwerten weniger ins Gewicht fallen. Sie wird als Standardfehler des Mittelwerts bezeichnet und ist definiert für die Grundgesamtheit durch

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}. \tag{22}$$

Mit n als Anzahl der Daten der Grundgesamtheit und der Standardabweichung σ. Für eine Stichprobe mit der Standardabweichung s und der Datenmenge n ist

$$\sigma(\bar{x}) = \frac{s}{\sqrt{n}}. \tag{23}$$

Mit steigender Stichprobengröße wird der Standardfehler kleiner und nähert sich dem Mittelwert der Grundgesamtheit. Dieses Maß erlaubt jetzt eine Intervallabschätzung um den geschätzten Mittelwert. Es wird in Prozent vom vorgegebenen Vertrauensgrad ausgedrückt, mit dem der Mittelwert der Stichprobe dem Mittelwert der Grundgesamtheit angenähert sein soll. Der so definierte Wertebereich wird als Konfidenzintervall bezeichnet.

3.1 Konfidenzintervall für normalverteilte Stichproben

Teilt man die Differenz vom Mittelwert der Stichprobe zum unbekannten Mittelwert der Grundgesamtheit μ durch den Standardfehler, so erhält man den z-Wert.

$$z = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}. \tag{24}$$

Sämtliche Normalverteilungen können durch diese einfache z-Transformation überführt werden in die Standardnormalverteilung. Vorausgesetzt, es liegt eine Normalverteilung vor. Geometrisch betrachtet entspricht dies der flächentreuen Transformation der Glockenkurve $f(\mu,\sigma^2)$ zur Glockenkurve f(0, 1) (Bild 10).

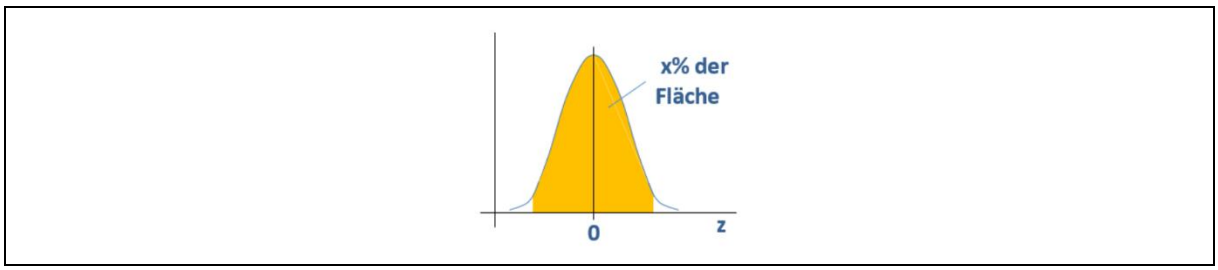

Bild 10. Standardnormalverteilung

Die Darstellung zeigt, dass der z-Wert sowohl negativ als auch positiv sein kann. Durch Umstellung von (24) ergibt sich das Konfidenzintervall aus

$$k_{\underset{\mu}{\sigma}}(\mu) = \bar{x} \pm z\frac{s}{\sqrt{n}}.. \tag{25}$$

Das Produkt aus z-Wert und Standardfehler heißt Fehlertoleranz der Stichprobe. Die Formel zeigt auch, dass sich oberer und unterer Konfidenzgrenzwert getrennt berechnen lassen.