Wilfried Rohm

Streuungsmaße bei Stichproben

- Mathematische / Fachliche Inhalte in Stichworten:

  Beschreibende Statistik, Standardabweichung, Varianz, Erwartungstreue Schätzung ,

- Kurzzusammenfassung

  Streuungsmaße sind in der beschreibenden Statistik von elementarer Bedeutung, jedoch erscheinen die verschiedenen Formelsätze oftmals verwirrend. Es soll hier sowohl ein mathematisches als auch intuitives Verständnis der verschiedenen Formeln für die Standardabweichung vermittelt werden. Zentral ist hierbei eine Simulation, welche zeigt, dass nur die Summe der quadrierten Abweichungen vom Mittelwert dividiert durch durch (n-1) statt n einen erwartungstreune Schätzwert für σ^2 einer Verteilung liefert.

- Lehrplanbezug (bzw. Gegenstand / Abteilung / Jahrgang):

  Angewandte Mathematik, 2.-5.Jahrgang, alle Abteilungen

- Mathcad-Version:

  erstellt mit Mathcad 11

Menüpunkte

- Verschiedene Streuungsmaße
- Erwartungstreue Schätzung der Standardabweichung (Simulation)
- Theoretische Begründung für die erwartungstreue Schätzung
- Die Standardabweichung des Mittelwertes - Zusammenhang zur Fehlerrechnung

Verschiedene Streuungsmaße bei Stichproben

Zum Menü

Zur Einführung ein bekanntes Gedicht:

```
Ein Mensch, der von Statistik hört,
denkt dabei nur an Mittelwert.
Er glaubt nicht dran und ist dagegen,
ein Beispiel soll es gleich belegen:

Ein Jäger auf der Entenjagd
hat einen ersten Schuß gewagt.
Der Schuß, zu hastig aus dem Rohr,
lag eine gute Handbreit vor.

Der zweite Schuß mit lautem Krach
lag eine gute Handbreit nach.
Der Jäger spricht ganz unbeschwert
voll Glauben an den Mittelwert:
Statistisch ist die Ente tot.

Doch wär´ er klug und nähme Schrot
- dies sie gesagt, ihn zu bekehren -
er würde seine Chancen mehren:
Der Schuß geht ab, die Ente stürzt,
weil Streuung ihr das Leben kürzt.
```

P.H.List, Professor für pharmazeutische Technologie in Marburg

Die üblichen Streuungsmaße sind:

- **Die Spannweite / der Range R:**

  $R = x_{max} - x_{min}$ Also: Maximalwert der Stichprobe - Minimalwert der Stichprobe

  Diese Größe ist nicht allzu beliebt, weil sie zu empfindlich auf Extremwerte reagiert und speziell bei vielen Stichprobenwerten das Problem besteht, dass einfach nur 2 Werte diese Kenngröße bestimmen.
  Daher wird diese Kenngröße bestenfalls als "Grundinformation" verwendet bzw. bei kleinen Stichprobengrößen - so wird etwa in der Qualitäts-Regelkartentechnik (wo n= 5 ein häufig verwendeter Stichprobenumfang ist) gerne mit dem Range statt mit der Standardabweichung gerechnet, da er leicht ohne Taschenrechner / Computer vor Ort von "einfachen Arbeitern" ermittelt werden kann.

- **Die mittlere lineare Abweichung**

  Das ist ein recht anschauliches Maß:

  $$\delta = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| x_i - \bar{x} \right|$$

  Es wird die durchschnittliche Abweichung vom arithmetischen Mittelwert berechnet - der Betrag ist notwendig, weil ohne ihn sich die positiven und negativen Abweichungen vom Mittelwert immer auf Null aufsummieren - der arithmetische Mittelwert wird ja gerade so definiert, dass die Summe aller Abweichungen gleich 0 sein muß.

- **Varianz und Standardabweichung**

  Trotz seiner Anschaulichkeit wird die mittlere lineare Abweichung als Streuungsmaß SELTEN verwendet. Der Grund liegt darin, dass seit C.F.Gauß Abweichungen (statt Beträge zu nehmen) nahezu grundsätzlich quadriert werden - ein ganzes Theoriegebäude (Fehlerrechnung, Wahrscheinlichkeitsrechnung, Statistik) baut gerade darauf auf. Das erklärt, warum gemeinhin das Kennmaß der Standardabweichung häufig mit dem Begriff "Streuung" einfach gleichgesetzt wird.

  Duch das Quadrieren wird ähnlich wie durch den Betrag die gegenseitige Auslöschung der Abweichungen verhindert.

Man erhält als eine mittlere quadratische Abweichung vom Mittelwert, also:

$$\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \mu\right)^2$$

Diese Größe wird VARIANZ genannt.

Aus Dimensionsgründen (die Abweichung soll ein LINEARES Maß sein), wird die Wurzel genommen und das Ergebnis STANDARDABWEICHUNG genannt:

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \mu\right)^2}$$

Wir haben in den obigen Formeln allerdings μ statt $\bar{x}$ stehen. Das hat einen guten Grund!
Die eben angegebenen Formeln gelten nur dann, wenn aus den Originaldaten (d,h, den Datenwerten einer vorliegenden GRUNDGESAMTHEIT bzw. "POPULATION") die Were für Varianz und Standardabweichung ermittelt werden. Die Größe μ deutet an, dass der Mittelwert dieser Grundgesamtheit gemeint ist!

In der Regel wird jedoch dieser Wert unbekannt sein und wird durch den arithmetischen Mittelwert $\bar{x}$ einer entsprechenden Stichprobe GESCHÄTZT. Dann lautet der richtige Formelsatz:

Stichprobenvarianz:

$$\boxed{s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

Stichprobenstandardabweichung

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

In diesem Artikel geht es im Weiteren nun darum, diesen kleinen (aber nicht unwesentlichen!) Unterschied im Formelsatz zu erklären.
Man kann dies auf verschiedenen Ebenen und verschiedenen "Exaktheitsniveaus" vornehmen.
Ich möchte hier verschiedene Erklärungen anbieten, von rein heuristisch über Simulationen bis zu einer mathematisch formulierten Begründung.

Heuristische Begründungen für die Wahl von (n-1) statt n:

Variante 1: Wird der "wahre Mittelwert" der Grundgesamtheit (μ) durch den arithmetischen Mittelwert einer Stichprobe GESCHÄTZT, ergibt sich eine gewisse Unsicherheit durch die Schätzung. Daher muss das Streuungsmaß etwas GRÖSSER ausfallen, weil weniger Information vorhanden ist. Dies leistet die Division durch einen kleineren Wert (n-1) statt n.
Dass gerade n-1 der "richtige" Wert ist, wurde damit allerdings noch nicht erläutert.

Variante 2: μ wurde durch $\bar{x}$ geschätzt. Für den Mittelwert gilt : $\bar{x} = \sum_{i=1}^{n} x_i$.

Das bedeutet: Die Messwerte x_i bestimmen den Mittelwert und (gemäß obiger Formeln) die Standardabweichung. Damit hier kein Widerspruch entstehen kann, sind von den n Messwerten höchstens n-1 "frei wählbar". Durch den vorgegebenen bzw. betreits zuvor berechneten Mittelwert ist der n-te Wert fixiert. Daher wird die größere Unsicherheit (von der in Variante 1 gesprochen wurde) durch die Division durch n-1 (dem "Freiheitsgrad f = n -1" der Stichprobe) formelmäßig umgesetzt (durchschnittliche quadratische Abweichung bezogen auf f = n-1 Werte!)

Erwartungstreue Schätzung der Standardabweichung (Simulation)

Zum Menü

Die folgende Simulation soll zeigen, dass s^2 einen sogenannten ERWARTUNGSTREUER Schätzwert für σ^2 darstellt, währenddessen bei einer Division durch n statt (n-1) der wahre Wert der Grundgesamtheit (σ^2) in der Regel bzw. tendenziell unterschätzt wird, sodass diese Schätzung nur als "assymptotisch erwartungstreu" (d.h. für n gegen Unendlich) zu bezeichnen ist.

Dies sieht man natürlich nur dann, wenn
- die Anzahl n (Stichprobenumfang) nicht zu groß ist
- die Anzahl der Simulationen genügend groß ist.

Um dies entsprechend zu demonstrieren, sollte man diese Größen in der folgenden Simulation variieren:

Versuchsmodell

Aus einer μ-σ-normalverteilten Grundgesamtheit werden N Stichproben zu je n Werten ("Meßwerte") entnommen. Von jeder Stichprobe wird die Stichprobenvarianz ermittelt und Stichprobenstandardabweichung ermittelt (auf 2 Arten: Division durch n-1 bzw. Division durch n). Anschließend werden beide Werte mit der "wahren" Standardabweichung σ graphisch verglichen!.

$\mu := 100$ $\sigma := 0.05$ Die Parameter der normalverteilten Grundgesamtheit

$n = 5$ n: Anzahl der Meßwerte je Simulation

$N = 10000$ N: Anzahl der Simulationen

n und N werden vor dem Diagramm unten bestimmt

$j := 0 .. N-1$

$x_j := (\mathrm{rnorm}(n, \mu, \sigma))$ Erzeugung von n normalverteilten Zufallszahlen ("Meßwerten") je Simulation

Nun werden die Stichprobenvarianzen s^2 bzw. sn^2 je Simulation berechnet und in einem Vektor abgelegt:

$$sq_j := \frac{1}{n-1} \cdot \left[\sum_{i=0}^{n-1} \left[\left(x_j\right)_i - \mathrm{mittelwert}\left(x_j\right) \right]^2 \right]$$

Erwartungstreuer Schätzwert für σ^2 !
(Wie unten gezeigt wird)

$$sqn_j := \frac{1}{n} \cdot \left[\sum_{i=0}^{n-1} \left[\left(x_j\right)_i - \mathrm{mittelwert}\left(x_j\right) \right]^2 \right]$$

Nicht erwartungstreuer, sondern bloß assymoptotisch erwartungstreuer Schätzwert für σ^2 !!!
(Wie unten gezeigt wird)

$$sq_{mittel} := \frac{1}{N} \cdot \sum_j sq_j \qquad s_{mittel} := \sqrt{sq_{mittel}}$$

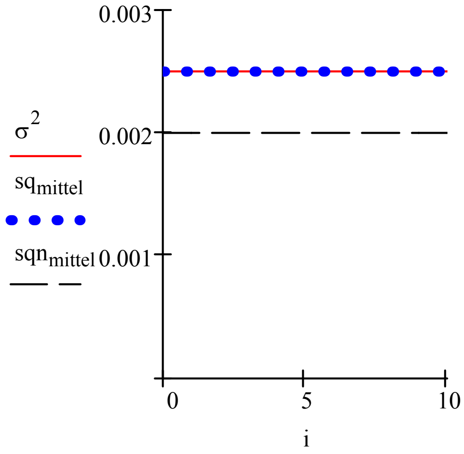
$$sqn_{mittel} := \frac{1}{N} \cdot \sum_j sqn_j \qquad sn_{mittel} := \sqrt{sqn_{mittel}}$$

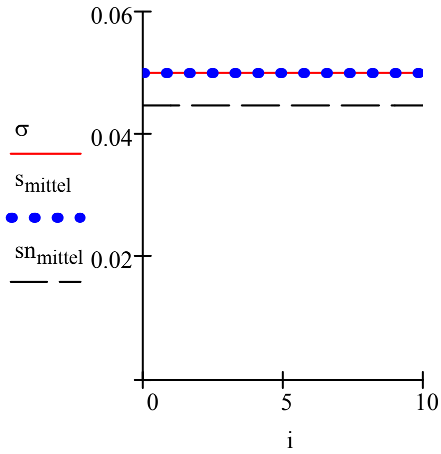
Berechnung der (bezogen auf die N Simulationen) durchschnittlichen Stichprobenvarianz bzw. Stichprobenstandardabweichung

$i := 0 .. 10$ Nur Hilfsgröße für die Grafik

$n \equiv 5$ Stichprobenumfang

$N \equiv 10000$ Anzahl der Simulationen





Beobachtungen :

Es wird die Wahl folgender Werte für n bzw. N empfohlen

$n \leq 10$ mit $N \geq 1000$

Das Ergebnis zeigt, dass hier σ bzw σ^2 der Grundgesamtheit bei der Division durch n deutlich unterschätzt wird. Die Division durch (n-1) hingegen führt optisch auf einen erwartungstreuen Schätzwert!

$n \geq 50$ oder $n \geq 100$ mit $N \geq 1000$

Das Ergebnis zeigt, dass hier der Unterschoied zwischen der Division durch n buzw. (n-1) optisch sehr klein ist bzw. erst bei einem entsprechenden Zoom sichtbar wird.
Da für größer werdendes n auch bei der Division durch n eine entsprechend gute Schätzung der Standarabweichung σ der Grundgesamtheit erfolgt, spricht man von einem asymptotisch erwartungstreuen Schätzwert.

$n \leq 10$ mit $N \leq 100$

Das Ergebnis zeigt, dass hier die Versuchszahl noch zu klein ist, um klare Aussagen treffen zu können.

Theoretische Begründung für die erwartungstreue Schätzung der Varianz der Grundgesamtheit:

Zum Menü

μ .... Mittelwert der Grundgesamtheit

$\bar{x}$ .... Mittelwert einer Stichprobe des Umfangs n mit Einzelwerten x_i

$$(x_i - \mu) = (x_i - \bar{x}) + (\bar{x} - \mu)$$

Trick, um $\bar{x}$ in die Formel der Standardabweichung zu bekommen

$$(x_i - \mu)^2 = (x_i - \bar{x})^2 + (\bar{x} - \mu)^2 + 2 \cdot (x_i - \bar{x}) \cdot (\bar{x} - \mu)$$

Nun erfolgt die Summation über alle n:

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n \cdot (\bar{x} - \mu)^2 + 2 \cdot (\bar{x} - \mu) \cdot \sum_{i=1}^{n} (x_i - \bar{x})$$

Wegen der Defintion des arithmetischen Mittelwertes gilt: $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

Damit und nach Division durch n folgt:

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

Daraus erkennt man, dass die n-Gewichtung **KEINEN erwartungstreuen Schätzwert** für die Varianz σ^2 der Grundgesamtheit liefert. Der "Fehler" stellt die quadrierte Abweichung der Stichprobenmittelwerte vom Gesamtmittelwert dar. Da gilt: $(\sigma_{mittelwert.})^2 = \frac{\sigma^2}{n}$ wird $(\bar{x} - \mu)^2$ durch $\frac{\sigma^2}{n}$ abgeschätzt.(Begründung im nächsten Abschnitt)

Daher :

$$\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{\sigma^2}{n} \qquad \text{bzw.} \qquad n \cdot \sigma^2 - \sigma^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Ab dieser Zeile ist jedoch das "=" - Zeichen als "ungefähr" zu interpretieren - was man gemeinhin dadurch andeutet, dass man σ durch s ersetzt. Also:

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{s^2}{n} \qquad \text{bzw.} \qquad n \cdot s^2 - s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

ERWARTUNGSTREUER Schätzwert für die Varianz der Grundgesamtheit
Da es sich um eine Schätzung handelt, wird σ durch s ersetzt!

Die Standardabweichung des Mittelwertes - Zusammenhang zur Fehlerrechnung

Zum Menü

Gauss´sches Fehlerfortpflanzungsgesetz

Ist $z = f(x,y,z)$ mit mittleren Absolutfehlern $\Delta x, \Delta y, \Delta z$

so gilt für den mittleren Absolutfehler Δz (ungefähr) (Fehlerrechnung!)

$$\Delta z = \sqrt{\left(\frac{d}{dx} f(x,y,z)\right)^2 \cdot \Delta x^2 + \left(\frac{d}{dy} f(x,y,z)\right)^2 \cdot \Delta y^2 + \left(\frac{d}{dz} f(x,y,z)\right)^2 \cdot \Delta z^2}$$

Wir berechnen nun nach dieser Formel den "mittleren Fehler des Mittelwertes" bei n mit gleicher Präzision durchgeführten Messungen

$$\bar{x} = f(x_1, x_2, ..., x_n) = \frac{x_1 + x_2 + .. + x_n}{n} = \frac{x_1}{n} + \frac{x_2}{n} + ... + \frac{x_n}{n}$$

Wir setzen: $\Delta x_i = \sigma$

Es gilt: $\frac{d}{dx_i} f(x_1, x_2, ..., x_n) = \frac{1}{n}$

Daher ist: $\Delta x_{mittelwert} = \sigma_{mittelwert} = \sqrt{\left(\frac{1}{n}\right)^2 \cdot \sigma^2 + \left(\frac{1}{n}\right)^2 \cdot \sigma^2 + ... + \left(\frac{1}{n}\right)^2 \cdot \sigma^2}$

Also : $\sigma_{mittelwert} = \sqrt{\frac{n \cdot \sigma^2}{n^2}} = \frac{\sigma}{\sqrt{n}}$

Dies ist das sogenannte "$\sqrt{n}$ -Gesetz", das sich aus dem Zentralen Grenzwertsatz der Mathematischen Statsitik ergibt.
Es wurde daher ein schöner Zusammenhang zwischen der Fehlerrechung und der Statistik aufgezeigt!

Zum Menü