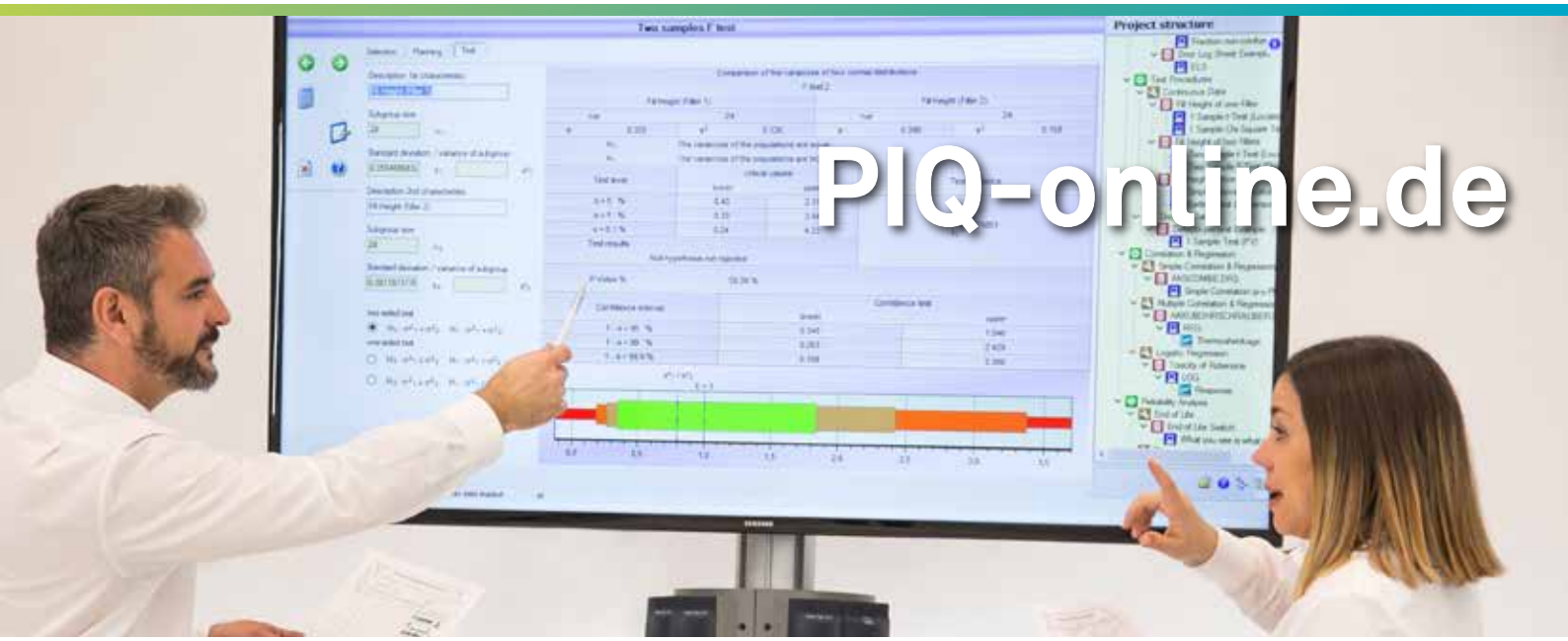
Die Lage der Dinge: Mittelwert oder Median?

Dipl.-Ing. Roman Wenig | Q-DAS GmbH



Dass der Median der bessere Schätzer für die Lage eines Qualitätsmerkmals ist, dürfte sich inzwischen herumgesprochen haben. Was sich noch nicht herumgesprochen zu haben scheint ist, was damit gemeint ist.

Gezeigt werden soll, welcher Stichprobenkennwert – Mittelwert oder Median – eine Datenreihe bezüglich ihrer mittleren Lage am besten repräsentiert. Dafür werden beispielhaft 1.000 Messwerte verwendet, die die Durchlaufzeit in einem Herstellprozess beschreiben.

Typisches Vorgehen: Grafische und numerische Analysen

Einen ersten Überblick bietet das Histogramm. Es soll die folgenden Fragen beantworten:

- Wie sind die Daten verteilt? Symmetrisch oder nicht symmetrisch? Ein- oder mehrgipflig?
- Wo befindet sich die mittlere Lage der Daten?
- Wie streuen die Daten?
- Sind alle Klassen besetzt – oder gibt es Lücken in den Daten?

Im Histogramm dargestellt, ergibt sich für die Durchlaufzeiten folgendes Bild:

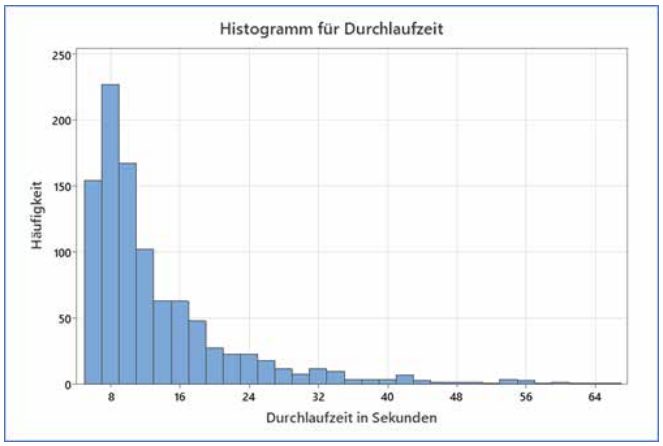


Bild 1: Histogramm der Durchlaufzeiten

Zu erkennen ist:

- Es liegt eine linkssteile (rechtsschiefe), eingipflige Verteilung vor. Diese Verteilung ist typisch für Zeiten: Schneller als ein bestimmter „natürlicher" unterer Grenzwert geht es meistens nicht, langsamer jedoch immer.
- Es wurden Durchlaufzeiten im einstelligen Sekundenbereich bis hin zu etwas über einer Minute beobachtet.
- Die meisten Werte wurden bei etwa 8 Sekunden beobachtet: Dort befindet sich der Gipfel.
- Alle Klassen sind mit Daten besetzt.

Mit etwas Erfahrung kann ausgesagt werden, dass in der Darstellung der Durchlaufzeiten im Histogramm keine Auffälligkeiten zu erkennen sind.

Eine Frage, die das obige Histogramm jedoch nicht beantworten kann, könnte nun sein: Wie groß ist die mittlere Durchlaufzeit? Anders gefragt:

- Welcher Zahlenwert repräsentiert alle Durchlaufzeiten am besten?
- Von welchem Zahlenwert haben alle Durchlaufzeiten die geringste Abweichung?

Zum besseren Verständnis: „Mittlere Durchlaufzeit" bedeutet, dass ein Kennwert gesucht wird, der die Natur des Qualitätsmerkmals mit einem Zahlenwert ausdrückt. Die gesamte beobachtete Variabilität wird also auf einen Zahlenwert reduziert.

Prinzipiell kommt für die Beschreibung der mittleren Lage jeder beobachtete Wert infrage: Die Wahrheit wird also zwischen dem Kleinstwert (5,21 s) und dem Größtwert (65,71 s) zu finden sein. Dass die beiden Extremwerte jedoch wenig geeignet sind, liegt auf der Hand.

Die typischerweise verwendeten statistischen Kenngrößen zur Beschreibung der mittleren Lage eines Qualitätsmerkmals sind der arithmetische Mittelwert und der Median. Ohne auf die exakte Ermittlung eingehen zu wollen, betragen sie für die Durchlaufzeiten:

- arithmetischer Mittelwert: 13,75 s
- Median: 10,49 s

Werden die beiden Werte dem Histogramm hinzugefügt, ergibt sich folgendes Bild:

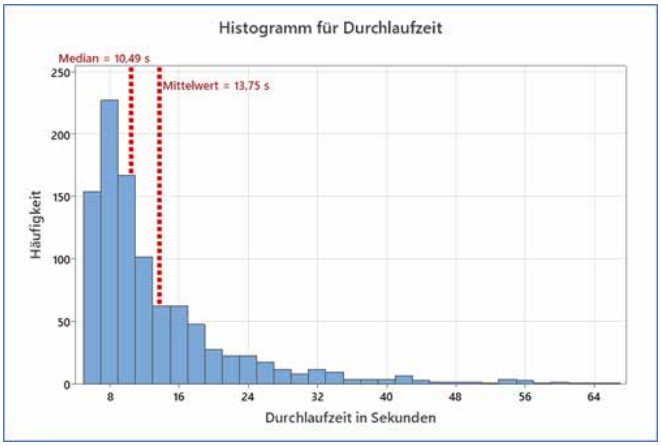


Bild 2: Histogramm der Durchlaufzeiten mit ergänztem Mittelwert und Median

Diese Darstellung scheint immer noch wenig geeignet zu sein, denjenigen Zahlenwert zu identifizieren, der alle Durchlaufzeiten am besten repräsentiert.

Was bedeutet eigentlich „am besten" hinsichtlich aller Durchlaufzeiten? Der gesuchte Zahlenwert soll die geringsten Abweichungen zu jeder Durchlaufzeit hervorrufen, möglichst gar keine, also null.

Diese Aussage kann dazu genutzt werden, um zwei neue Spalten zu erzeugen:

- x – xquer: Von jeder beobachteten Durchlaufzeit wird der Mittelwert abgezogen.
- x – median: Von jeder beobachteten Durchlaufzeit wird der Median abgezogen.

Durchlaufzeit	x-xquer	x-median
5,5108	-8,2422	-4,9752
5,5176	-8,2354	-4,9684
5,5423	-8,2107	-4,9437
5,5884	-8,1646	-4,8976
5,6164	-8,1366	-4,8696

Tabelle 1: Struktur der Daten: Durchlaufzeit und zwei „Abweichungsspalten“

Damit kann die folgende Aussage getroffen werden: Diejenige Spalte, die die geringsten Abweichungen enthält, «gewinnt». Das bedeutet: Der Stichprobenkennwert, mit dem die kleinsten Abweichungen bezüglich aller Durchlaufzeiten erzeugt werden, repräsentiert die mittlere Durchlaufzeit am besten.

Um die Abweichungen grafisch darstellen zu können, werden die beiden neu erzeugten Spalten aufsteigend (der Größe nach) sortiert. Beide sortierten Spalten werden sodann gemeinsam in einem Flächendiagramm dargestellt.

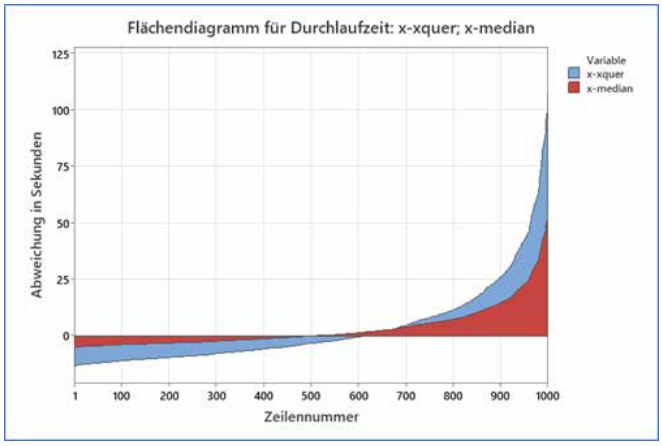


Bild 3: Darstellung der sortierten Abweichungen im Flächendiagramm

Unschwer ist zu erkennen:

- Zwei farblich unterschiedliche Flächen gruppieren sich um die null herum.
- Die Fläche „x-median“ liegt dichter an der null, aber
- es gibt einen Bereich, in dem die Flächen übereinander liegen.
- Es gibt zwei Punkte, an denen die Flächen die null schneiden.

Wenn dieses interessante Intervall vergrößert dargestellt wird, ergibt sich folgendes Bild:

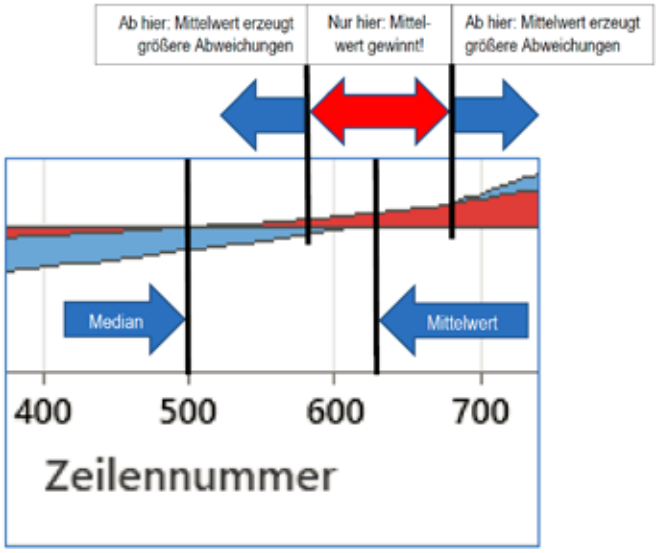


Bild 4: Ausschnitt aus Bild 3, vergrößerte Darstellung

Basierend auf der grafischen Darstellung kann geschlussfolgert werden:

- Der Median führt bei den meisten Durchlaufzeiten zu geringeren Abweichungen als der Mittelwert.
- Es gibt ein kleines Intervall, in dem der Mittelwert die Durchlaufzeiten besser als der Median repräsentiert. Dieses Intervall umfasst im gewählten Datensatz etwa 10 % der Durchlaufzeiten.

Zusammenfassung

Was wie Zahlenspielerei anmutet, hat durchaus praktische Konsequenzen. Wenn der Mittelwert anstelle des Medians als Schätzer der mittleren Durchlaufzeit verwendet wird, führt das:

- infolge des Zahlenwertes des Mittelwertes, der größer ist als der des Medians
  - in der Planung zu Vorgabezeiten, die zu lang sind, damit
  - in der Produktion zu weniger hergestellten Teilen, damit
  - zu weniger Umsatz.
- infolge der Abweichungen, die größer als beim Median sind:
  - zur Erhöhung der Durchlaufzeiten bis hin zu
  - Verzögerungen in der Lieferung.

Wenn es also um die Lage der Dinge geht, ist der Median meistens der bessere Schätzer. Das darf auch bei symmetrischen Verteilungen in Erwägung gezogen werden.

Haben wir Ihr Interesse geweckt? Bitte richten Sie Fragen direkt an den Autor:

Q-DAS GmbH
Reichenhainer Str. 29a
09126 Chemnitz
www.q-das.com
roman.wenig@hexagon.com