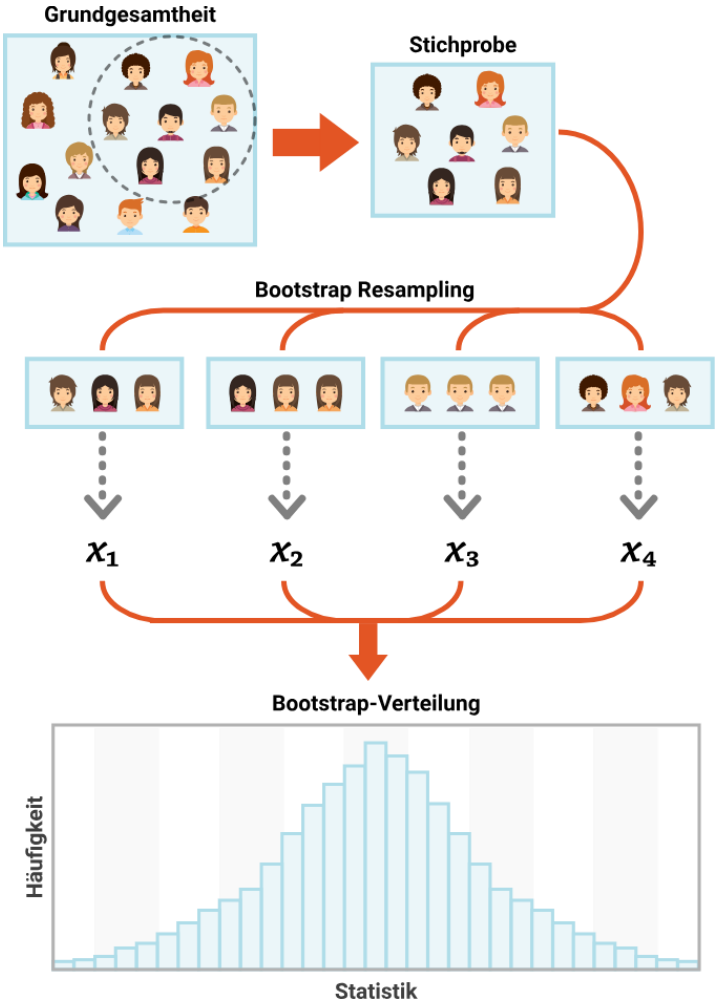
Was ist Bootstrapping?

Eine Stichprobenverteilung beschreibt die Wahrscheinlichkeit, mit der jeder mögliche Wert einer Statistik aus einer Zufallsstichprobe einer Grundgesamtheit erhalten wird, oder mit anderen Worten, welcher Anteil von allen Zufallsstichproben mit dem betreffenden Stichprobenumfang den betreffenden Wert ergibt.

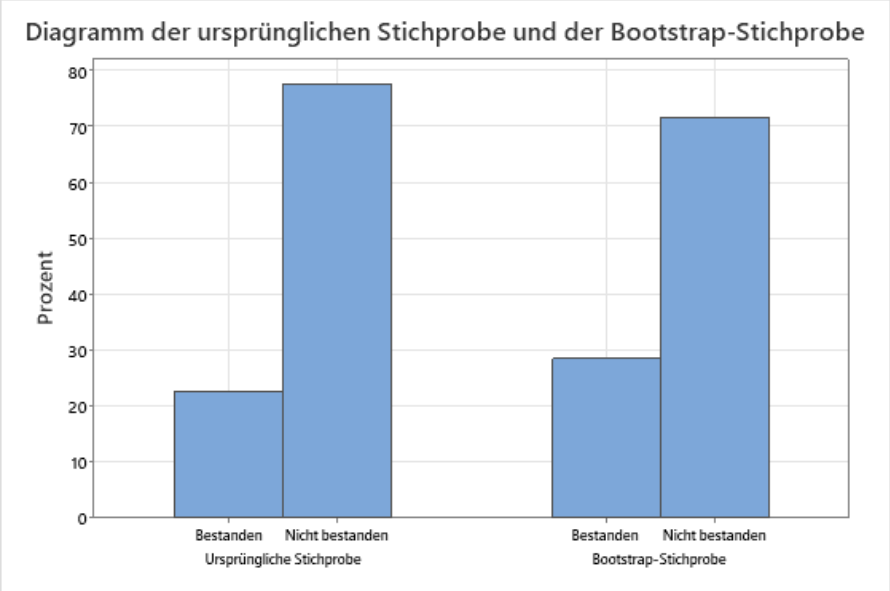
Bootstrapping gehört zu einer größeren Klasse von Verfahren, die empirische Verteilungen durch die erneute Stichprobenziehung aus dem ursprünglichen Datensatz aufstellen, die Resampling Methoden. Wichtig ist, das meist mit Zurücklegen gezogen wird, d.h. in einer Stichprobe kann ein Wert mehr als einmal vorkommen, wie in der Abbildung unten.

Bootstrapping ist ein Verfahren zum Schätzen der Stichprobenverteilung, bei dem mehrere Stichproben mit Zurücklegen aus einer einzigen Zufallsstichprobe gezogen werden. Diese wiederholt gezogenen Stichproben werden als Stichprobenwiederholungen bezeichnet. Jede Stichprobenwiederholung hat den gleichen Stichprobenumfang wie die ursprüngliche Stichprobe.

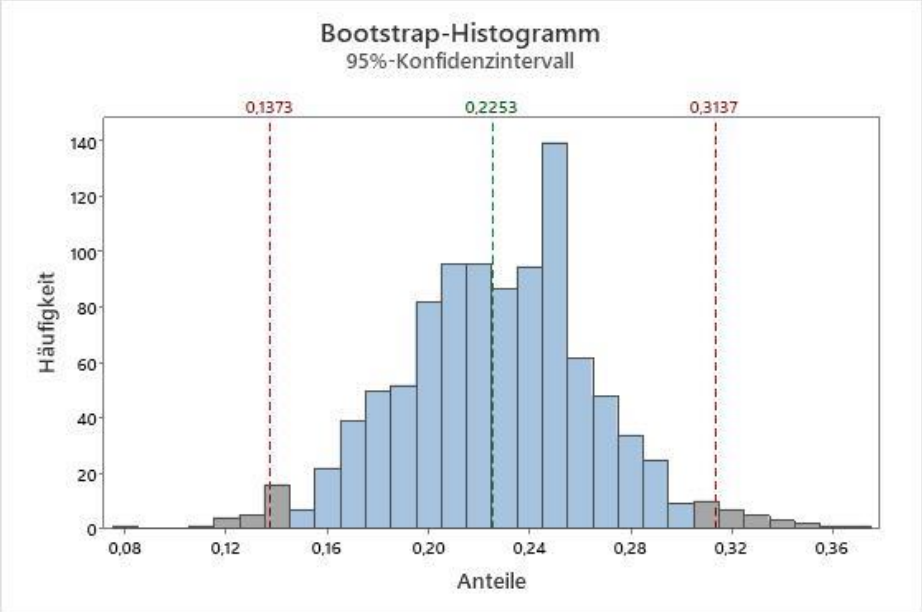
Die ursprüngliche Stichprobe stellt die Grundgesamtheit dar, aus der sie gezogen wurde. Daher stellen die Stichprobenwiederholungen aus dieser ursprünglichen Stichprobe das Ergebnis dar, das beim Ziehen vieler Stichproben aus der Grundgesamtheit erhalten würde. Die Bootstrap-Verteilung einer Statistik, die auf den Stichprobenwiederholungen basiert, stellt die Stichprobenverteilung der Statistik dar.



Angenommen, Sie möchten beispielsweise die Stichprobenverteilung des Anteils von blauen M&Ms schätzen. Sie öffnen ein zufällig ausgewähltes Päckchen und stellen fest, dass es 102 M&Ms enthält, von denen 23 (22,5 %) blau sind. Durch die wiederholte Stichprobennahme mit Zurücklegen aus dieser ursprünglichen Stichprobe wird die mögliche Zusammensetzung der Grundgesamtheit abgebildet. Zum Erhalten einer Stichprobenwiederholung wird ein M&M zufällig aus der ursprünglichen Stichprobe ausgewählt, die Farbe wird notiert, und das M&M wird in die Stichprobe zurückgelegt. Dieser Vorgang wird 102 Mal (Umfang der ursprünglichen Stichprobe) wiederholt, um eine einzige Stichprobenwiederholung zu vervollständigen. Im folgenden Balkendiagramm wird eine einzige Bootstrap-Stichprobe dargestellt, die aus der ursprünglichen Stichprobe gezogen wurde.



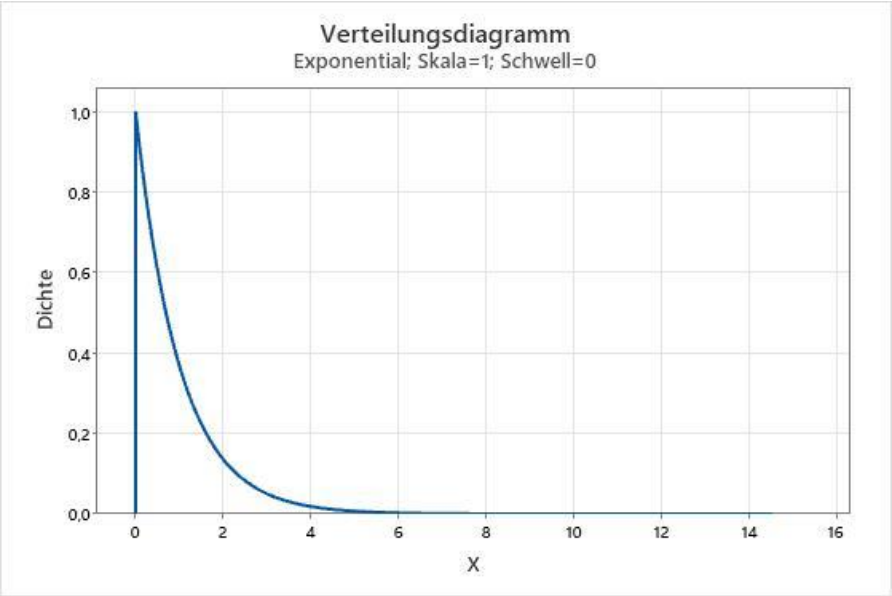
Da die Stichprobenwiederholung durch Stichprobennahme mit Zurücklegen erfolgt, entspricht der Anteil der Bootstrap-Stichprobe im Allgemeinen nicht genau dem ursprünglichen Anteil. Dieses Balkendiagramm zeigt, dass in der ursprünglichen Stichprobe festgestellt wurde, dass ca. 22,5 % der M&Ms blau sind, während die Bootstrap-Stichprobe ergab, dass etwa 28,4 % der M&Ms blau sind. Um eine Bootstrap-Verteilung zu erstellen, ziehen Sie viele Stichprobenwiederholungen. Im folgenden Histogramm wird die Bootstrap-Verteilung für 1000 Stichprobenwiederholungen des ursprünglichen M&M-Päckchens dargestellt.



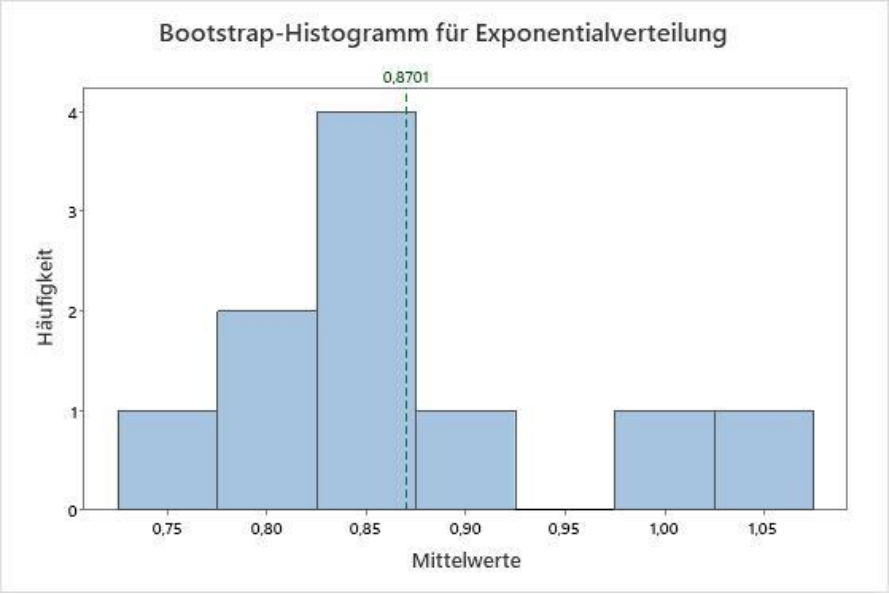
Das Zentrum der Bootstrap-Verteilung befindet sich bei annähernd 22,5 %; hierbei handelt es sich um einen Schätzwert des Anteils der Grundgesamtheit. Die roten Referenzlinien stellen ein 95%-Konfidenzintervall dar. Die mittleren 95 % der Werte aus der Bootstrap-Verteilung liefern ein 95%-Konfidenzintervall für den Anteil der Grundgesamtheit von blauen M&Ms. In diesem Beispiel können Sie sich zu 95 % sicher sein, dass der Anteil von blauen M&Ms der Grundgesamtheit zwischen ungefähr 13,7 % und 31,4 % liegt.

Bootstrapping und der zentrale Grenzwertsatz

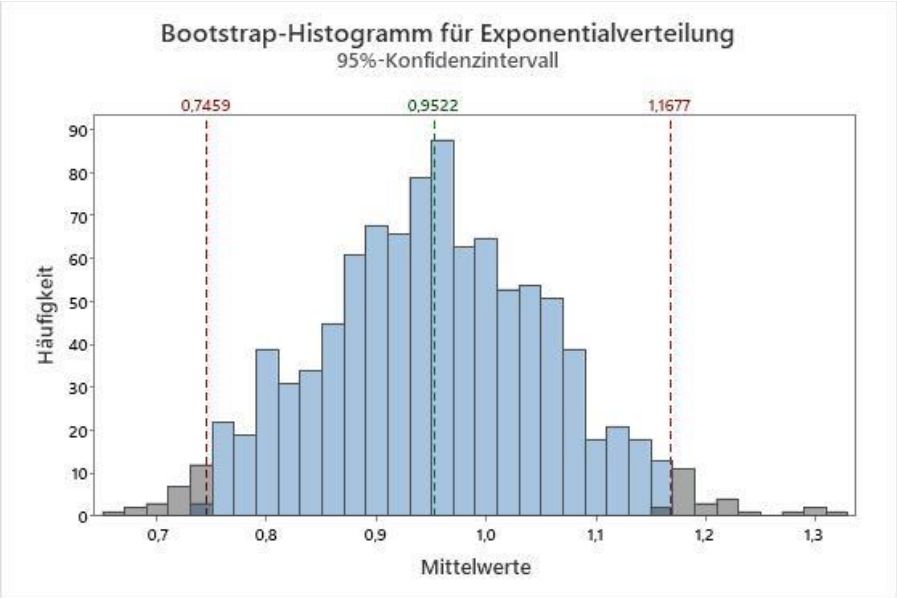
Der zentrale Grenzwertsatz ist ein grundlegender Satz der Wahrscheinlichkeitsrechnung und Statistik. Er besagt, dass die Verteilung von Zufallsvariablen, wobei es sich um den Mittelwert einer zufälligen Stichprobe aus einer Grundgesamtheit mit endlicher Varianz handelt, bei einem großem Stichprobenumfang ungefähr normalverteilt ist, und zwar unabhängig von der Verteilung der Grundgesamtheit. Durch Bootstrapping wird das Wesen des zentralen Grenzwertsatzes leichter verständlich. Betrachten Sie Daten, die aus einer Exponentialverteilung stammen.



Es ist deutlich ersichtlich, dass die Daten nicht normalverteilt sind. Nun ziehen wir jedoch eine Stichprobe von 50 Beobachtungen und erstellen eine Bootstrap-Verteilung der Mittelwerte von 10 Stichprobenwiederholungen.



Die Verteilung der Mittelwerte unterscheidet sich stark von der Exponentialverteilung. Sie ähnelt vielmehr einer Normalverteilung. Dieser Anschein verstärkt sich noch, wenn die Anzahl der Stichprobenwiederholungen vergrößert wird. Bei 1000 Stichprobenwiederholungen sieht die Verteilung der Mittelwerte der Stichprobenwiederholungen annähernd normalverteilt aus.



Quelle:

https://support.minitab.com/

https://statistikguru.de/