https://datascience-project.eu/

Training Fiche Vorlage

Titel	Stichprobentheorie
Schlüsselwörter (Meta-Tags)	Datenerhebung, statistische Schlussfolgerungen, Schätzung, Bestimmung des Stichprobenumfangs, einfache Zufallsstichproben, geschichtete Stichproben
Sprache	Deutsch
Zielsetzungen / Ziele / Lernergebnisse	**Ziel dieses Moduls ist es, die Grundlagen der Stichprobentheorie einzuführen und zu erläutern.** **Dieses Modul zeigt dir, wie du:** - die Unterschiede zwischen Populationen und Stichproben erkennst - die am häufigsten angewandten Stichprobenverfahren verwenden kannst - die optimale Stichprobengrößen findest

Lehrgang:	
Datenwissenschaftliche Kompetenz	
Modul Datenvisualisierung und visuelle Analyse	X
Einführung in die Datenwissenschaft für Human- und Sozialwissenschaften	
Datenwissenschaft für den guten Zweck	
Datenjournalismus und Geschichtenerzählen	

Beschreibung	In diesem Schulungsmodul werden wir die Grundlagen der Stichprobentheorie lernen. Im Zusammenhang mit der Theorie der statistischen Inferenz, genauer gesagt mit den Instrumenten, die die Berechnung von Konfidenzintervallen ermöglichen, werden wir die Verfahren untersuchen, die zur Ermittlung des optimalen Stichprobenumfangs in Abhängigkeit von dem zu schätzenden Merkmal und der verwendeten Stichprobentechnik verwendet werden. In diesem Modul werden wir die Unterschiede zwischen stichprobenbasierten Daten und populationsbasierten Daten sowie die am häufigsten angewandten Stichprobenverfahren untersuchen: einfache und geschichtete Stichproben. Darüber hinaus werden wir die Regeln für die Ermittlung des optimalen Stichprobenumfangs unter Berücksichtigung einiger Ziele in Bezug auf das Vertrauen und die Fehlerspanne, die wir bei unseren Schlussfolgerungen haben möchten, untersuchen.

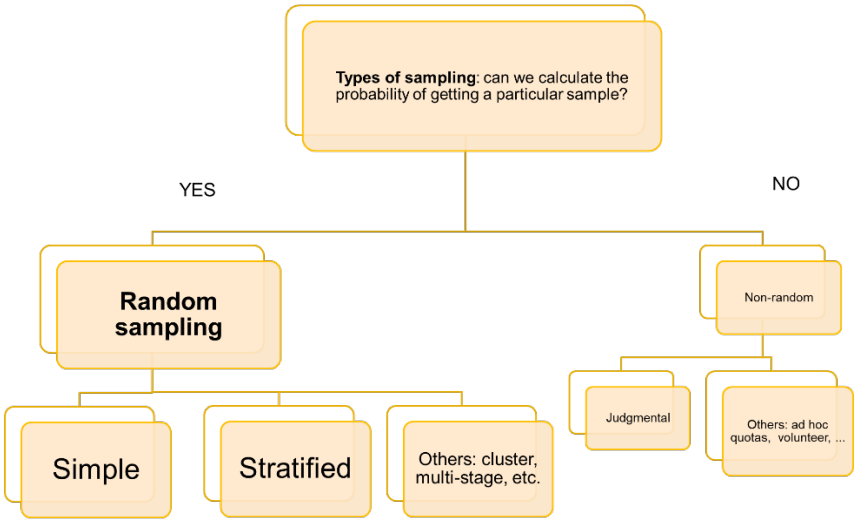
Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Inhalt in 3 Ebenen gegliedert	**1. EINLEITUNG** In der statistischen Analyse ist eine Grundgesamtheit (Population) eine Gruppe, für die wir einen Datensatz erstellen und einige Schlussfolgerungen ziehen wollen. Eine Erhebung ist ein Verfahren, mit dem wir die zu analysierenden Daten erhalten. Erhebungen können auf der Gesamtbevölkerung basieren (zensus- oder bevölkerungsbasiert), oder wir möchten eine repräsentative Teilmenge dieser Population auswählen. Diese Teilmenge wird als "Stichprobe" bezeichnet, wenn ihre Struktur dieselbe ist wie die der Grundgesamtheit. Daten aus Erhebungen, die eine Stichprobe heranziehen, werden als stichprobenbasiert bezeichnet. Warum werden Datensätze in Form von Stichproben erhoben, anstatt die gesamte Bevölkerung zu untersuchen (Erhebungen auf der Grundlage von Volkszählungen)? Letztere sind für Zählungen und erschöpfende Untersuchungen notwendig, erfordern jedoch den Einsatz umfangreicher Ressourcen, was zu hohen Kosten führt. Im Gegensatz dazu sind stichprobenartige Erhebungen geeignet, wenn die Bevölkerung homogen ist, da sie ein gutes Abbild der Bevölkerung darstellen. Außerdem sind sie die einzige Möglichkeit, wenn die Grundgesamtheit unendlich groß ist oder der Erhebungsprozess zu Informationszerstörungen führen kann. In jedem Fall sparen Stichproben Zeit und Kosten. In der Praxis haben wir in der Regel nicht die Ressourcen, um bevölkerungsbezogene Studien durchzuführen, so dass die Alternative darin besteht, unsere Analyse auf Stichproben zu stützen. Wenn wir unsere Schlussfolgerungen auf Stichprobendaten stützen, bedeutet dies, dass es eine inhärente Fehlerspanne gibt, auf die sich mehrere Faktoren auswirken können. Die Fehlermarge wird im Wesentlichen von drei Faktoren abhängen: a. Wie homogen die Daten in der Grundgesamtheit sind: Je heterogener die Daten sind, desto größer ist die Fehlermarge, wenn alle anderen Faktoren gleich bleiben. b. Der Stichprobenumfang: Je kleiner der Umfang, desto größer ist die Fehlermarge, wenn alles andere gleich bleibt. c. Das Stichprobenverfahren: abhängig von den Merkmalen Ihrer Daten. Bei (a) können wir nicht viel tun, aber bei den Punkten (b) und (c) gibt es einen gewissen Handlungsspielraum. Zu Punkt (c) über die angewandte Stichprobentechnik ist anzumerken, dass es eine große Vielfalt an verfügbaren Stichprobentechniken gibt, die wir anwenden können. Das nachstehende Diagramm zeigt diese Vielfalt in visueller Form:

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.



Wir können die Fehlermarge unserer Schlussfolgerungen nur kontrollieren, wenn wir mit Zufallsstichproben arbeiten, und die am häufigsten verwendeten Stichprobenverfahren sind die einfache Zufallsstichprobe und die geschichtete Zufallsstichprobe.

2. PROBENAHMETECHNIKEN

2.1. Einfache Zufallsstichprobe

Die einfache Zufallsstichprobe ist das elementarste Stichprobenverfahren, das auf einer Zufallsauswahl der untersuchten Beobachtungen beruht. Sie besteht darin, ausgehend von einer Auflistung der Einheiten der Grundgesamtheit, n dieser Einheiten zufällig auszuwählen. Aber selbst bei dieser einfachen Technik können einige Besonderheiten des Zufallsauswahlverfahrens festgelegt werden. So kann beispielsweise entschieden werden, ob die Stichprobe mit oder ohne Zurücklegen durchgeführt werden soll. Wird die Stichprobe mit Zurücklegen durchgeführt, bedeutet dies, dass jede Einheit, die nach dem Zufallsprinzip als Teil der Stichprobe ausgewählt wurde, nach jeder Ziehung wieder in die Grundgesamtheit aufgenommen wird. Dies bedeutet natürlich, dass eine Einheit mehr als einmal in die Stichprobe aufgenommen werden kann, aber es garantiert, dass die Bedingungen, unter denen die einzelnen Ziehungen stattfinden, gleich und konstant sind und die Ergebnisse der einzelnen Ziehungen unabhängig voneinander sind.

Im Gegensatz dazu wird bei einer einfachen Zufallsstichprobe ohne Zurücklegen jede Einheit nur einmal beprobt, aber wir können nicht garantieren, dass die Bedingungen entlang der Auswahlziehungen konstant sind. Stichproben mit und ohne Zurücklegen können bei kleinen Populationen zu sehr unterschiedlichen Ergebnissen führen. Sie sind nur dann gleichwertig, wenn die Größe der Grundgesamtheit (N) sehr groß ist.

Co-funded by the Erasmus+ Programme of the European Union
Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

2.2. Geschichtete Zufallsstichprobe

In manchen Fällen werden die Beobachtungen natürlich auf der Grundlage gemeinsamer Merkmale gruppiert. So werden beispielsweise Daten über die Lohnverteilung nach dem Wirtschaftszweig der Arbeitnehmer, ihrem Geschlecht oder ihrer Wohnregion gruppiert. Schichten werden als Teile der relevanten Population definiert, die eine hohe interne Homogenität aufweisen, auch wenn zwischen den Schichten eine große Variabilität besteht. Die geschichtete Stichprobe nutzt diese Gruppierung der Beobachtungen und wählt nach dem Zufallsprinzip eine Anzahl von Einheiten in jeder Schicht L (n_L) aus, so dass sich der Gesamtstichprobenumfang durch Addition der in jeder Schicht beprobten Elemente ergibt. Es gibt mehrere Kriterien für die Aufteilung des Gesamtstichprobenumfangs auf die Schichten, von denen die folgenden die häufigsten sind:

- Disproportional: gleiche Stichprobengröße in jeder Schicht
- Proportional: der Anteil der Stichprobenmitglieder entspricht dem Anteil der Bevölkerungsmitglieder in jeder Schicht
- Optimal: proportional zur Größe und Heterogenität (Varianz) der einzelnen Schichten

Unter den gleichen Bedingungen und mit den gleichen Anforderungen an Präzision und Konfidenzintervall können wir bestätigen, dass geschichtete Stichproben im Allgemeinen einen geringeren Stichprobenumfang erfordern als einfache Stichproben. Auf Fragen im Zusammenhang mit der Berechnung des Stichprobenumfangs wird im nächsten Punkt eingegangen.

3. BERECHNUNG DES OPTIMALEN STICHPROBENUMFANGS

Die goldene Regel für den Zusammenhang zwischen dem Stichprobenumfang und der Genauigkeit unserer Schätzungen lautet, dass die Fehlermarge umso geringer ist, je größer der Stichprobenumfang ist, wenn alle anderen Faktoren gleich bleiben. Die Beschaffung statistischer Daten, selbst wenn sie in Form einer Stichprobe erfolgt, kann jedoch kostspielig sein, und manchmal fehlen uns die Mittel für große Stichproben. Folglich gibt es eine Kompromisslösung, die den optimalen (minimalen) Stichprobenumfang festlegt, den wir angesichts unserer Anforderungen an die Genauigkeit (Fehlermarge) und das Vertrauen in unsere Schätzungen sowie die Heterogenität (Varianz) der relevanten Variablen in der Grundgesamtheit benötigen.

3.1 Lösung für eine einfache Stichproboben

Nehmen wir zunächst an, dass wir mit unserer Stichprobe den Mittelwert der Grundgesamtheit für eine kontinuierliche Variable schätzen wollen und dass unsere Stichprobe anhand einer einfachen Zufallsstichprobe ausgewählt wird. Die Formeln, die wir anwenden müssen, sind die folgenden:

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

$$n^* = k^2 \frac{\sigma^2}{e^2}$$

$$n = \frac{n^*}{1 + \frac{n^*}{N}}$$

Die Konstante k stammt aus einer Normalverteilung und wird größer, wenn wir das gewünschte Konfidenzintervall erhöhen, und das Symbol e steht für die Fehlermarge, die wir anzunehmen bereit sind. Zusätzlich müssen wir eine Annahme über die Homogenität der Variablen in der Grundgesamtheit treffen. Dies bedeutet, dass wir einen realistischen Wert (der in der Regel aus einer früheren Studie stammt) für die Varianz der Population σ^2 festlegen müssen.

In diesen Gleichungen ist n^* die Lösung für eine einfache Zufallsstichprobe mit Zurücklegen, n ist die Lösung für eine einfache Zufallsstichprobe ohne Zurücklegen und N ist die Größe der Grundgesamtheit. Generell gilt $n^* \geq n$, und beide Lösungen konvergieren, wenn N sehr groß ist.

Wenn wir daran interessiert sind, den Anteil (P) der Einheiten in einer Grundgesamtheit zu schätzen, die ein bestimmtes Merkmal aufweisen, lauten die Ausdrücke, die erforderlich sind, um den optimalen Stichprobenumfang für dieses Stichprobenverfahren zu finden, wie folgt:

$$n^* = k^2 \frac{P * (1 - P)}{e^2}$$

$$n = \frac{n^*}{1 + \frac{n^*}{N}}$$

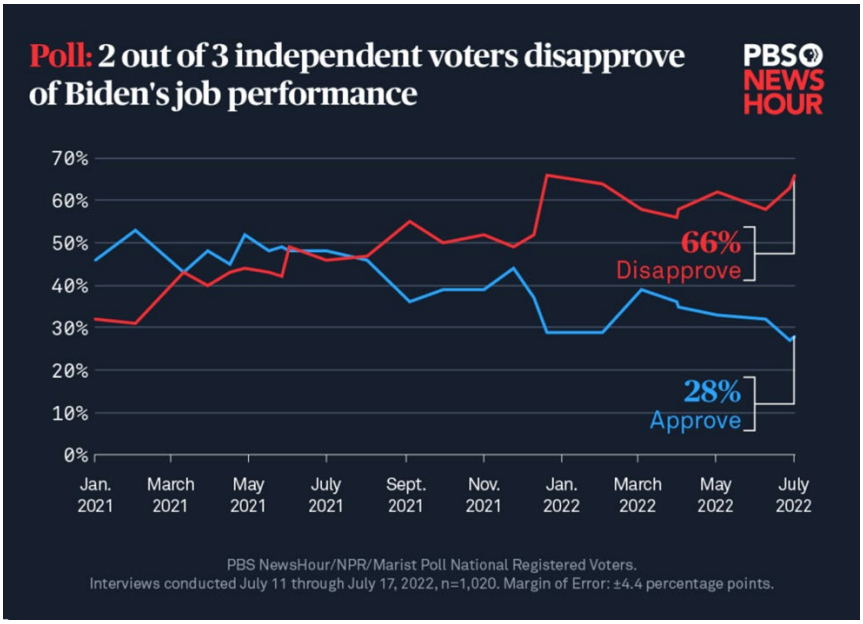
Auch hier stammt die Konstante k aus einer Normalverteilung und wird größer, wenn wir das gewünschte Konfidenzintervall erhöhen, und der Term e steht für die Fehlermarge, die wir bereit sind, anzunehmen. In diesem Fall müssen wir eine Annahme über den Wert von P*(1-P) machen, der die Varianz einer binären (ja/nein) Variable ist. Die übliche Lösung besteht darin, P=1-P=0,5 anzunehmen, so dass P*(1-P)=0,25 seinen Maximalwert annimmt.

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Wir können diese Technik anhand eines praktischen Beispiels veranschaulichen, wie Stichprobengrößen bestimmt werden und wie die Anwendung von R uns dabei helfen kann:

Der Public Broadcasting Service (PBS) in den USA schätzt regelmäßig den Prozentsatz der Bürger:innen, welche die Arbeit des US-amerikanischen Präsidenten gutheißen oder ablehnen. Im Fall von US-Präsident Joe Biden werden diese Umfragen seit Januar 2021 durchgeführt. Die folgende Grafik zeigt die Entwicklung der Schätzungen:



Bei einer kürzlich durchgeführten Umfrage in dieser Reihe wollte PBS Schätzungen mit einem Konfidenzniveau von 99 % erhalten, war bereit, eine Fehlermarge von ±4,4 % in Kauf zu nehmen, ging vom ungünstigsten Fall aus (übliche Lösung) und nahm an, dass der Prozentsatz der Befürworter (P) gleich dem Prozentsatz der Nichtbefürworter (1-P) ist. Wie viele Bürger:innen müssten unter diesen Bedingungen in die Stichprobe aufgenommen werden? Die oben dargestellten Gleichungen können in der Sprache R implementiert werden, um eine Lösung zu finden.

Zunächst müssen wir die erforderlichen Pakete installieren und laden:

```
#install and call the required package
install.packages("samplingbook")
library("samplingbook")
```

Später können wir diesen optimalen Stichprobenumfang durch Aufruf der Funktion "sample.size.prop" im Paket ermitteln. Diese Funktion ermöglicht eine Stichprobenziehung mit oder ohne Zurücklegen, obwohl es in der Praxis keine Unterschiede zwischen den Lösungen dieser beiden Alternativen gibt, da die Grundgesamtheit (N), aus der die Stichproben gezogen werden, sehr groß ist (wir

Co-funded by the Erasmus+ Programme of the European Union
Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

können willkürlich annehmen, dass N=200.000.000). Die folgenden Codeteile berechnen die jeweiligen Lösungen für eine Stichprobe ohne und mit Zurücklegen:

```
#calculation of simple random sample for estimating a population proportion
#the margin of error is "e" , the pop. proportion is assumed to be "P"
sample.size.prop(e=0.04,P=0.5,N=200000000,level = 0.99) #without replacement#
sample.size.prop(e=0.04,P=0.5,level = 0.99)       #with replacement#
```

Die in beiden Fällen als Lösung eine Stichprobengröße von etwa 1.000 Einheiten findet.

3.2. Lösung für geschichtete Stichproben

In diesem Abschnitt werden die Formeln für die Berechnung des Stichprobenumfangs bei geschichteten Stichproben ausführlich erläutert. Der Einfachheit und Klarheit halber konzentrieren wir uns nur auf den Fall der Schätzung eines Populationsmittelwerts und bieten die beiden häufigsten Lösungen an, die den Fällen der proportionalen (1) und der optimalen Verteilung (2) entsprechen:

$$(1) \quad n = \frac{\sum_{j=1}^{L} N_j \sigma_j^2}{N\frac{e^2}{k^2} + \frac{\sum_{j=1}^{L} N_j \sigma_j^2}{N}}$$

$$(2) \quad n = \frac{\frac{1}{N}\left(\sum_{j=1}^{L} N_j \sigma_j\right)^2}{N\frac{e^2}{k^2} + \frac{\sum_{j=1}^{L} N_j \sigma_j^2}{N}}$$

Wie bereits erwähnt, entspricht die Formel in beiden Fällen der Schätzung des Populationsmittelwerts für eine kontinuierliche Variable mit einer geschichteten Stichprobe ohne Zurücklegen. In diesen Ausdrücken steht N_j für die Größe der Schicht j und σ_j^2 für die Varianz der Variablen in dieser Schicht.

Ähnlich wie bei den Lösungen für einfache Zufallsstichproben können wir anhand eines praktischen Beispiels in der Sprache R veranschaulichen, wie der optimale Stichprobenumfang bei geschichteten Stichproben berechnet wird.

Angenommen, ein Wohlfahrtsverband führt eine Stichprobenerhebung durch, um die jährlichen Spenden seiner Mitglieder zu untersuchen, die je nach Alter in drei verschiedene Gruppen mit jeweils 100, 700 und 200 Mitgliedern eingeteilt sind. Aus einer Pilotstudie weiß die Wohltätigkeitsorganisation, dass die jeweiligen Standardabweichungen (σ_j) bei den jährlichen Spenden in jeder Gruppe 6 €, 30 € und 12 € betragen. Es soll der Mindeststichprobenumfang ermittelt werden, der erforderlich ist, um die mittlere jährliche Spende zu schätzen, wobei eine Fehlermarge von 2 € und ein Konfidenzniveau von 95 % angesetzt werden.

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Die folgenden Codezeilen berechnen den optimalen Stichprobenumfang und bieten die Lösungen für den Fall einer proportionalen und optimalen Verteilung, indem sie die Funktion "stratasize" aus dem Paket "samplingbook" in R aufrufen:

```
#============================================================================
#calculation of stratified random sample for estimating a population mean
#the margin of error is "e" , the pop. standard deviation is assumed to be "Sh"
#============================================================================

#proportional allocation
n_prop<-stratasize(e=2, level=0.95, Nh=c(100,700,200), Sh=c(6,30,12), type="prop")
#optimal allocation
n_opt<-stratasize(e=2, level=0.95, Nh=c(100,700,200), Sh=c(6,30,12), type="opt")

#display the results (already rounded up to the next integer)
n_prop
n_opt
```

Die entsprechenden Lösungen sind 390 und 339 Einheiten, wie im Folgenden beschrieben:

```
stratamean object: Stratified sample size determination

type of sample: prop

total sample size determinated: 390
> n_opt

stratamean object: Stratified sample size determination

type of sample: opt

total sample size determinated: 339
```

Schließlich können wir uns fragen, welcher dieser beiden Stichprobenumfänge auf die Schichten aufgeteilt werden soll. Dies kann durch den Aufruf der Funktion "stratasamp" im selben Paket erfolgen:

```
#============================================================================
#allocating the sample size
#============================================================================
# extract the sample size from the list
n_prop_int <- as.integer(n_prop$n)
n_opt_int <- as.integer(n_opt$n)

# allocate the sample size across strata: proportional allocation
stratasamp(n=n_prop_int, Nh=c(100,700,200), Sh=c(6,30,12), type="prop")

# allocate the sample size across strata: optimal allocation
stratasamp(n=n_opt_int, Nh=c(100,700,200), Sh=c(6,30,12), type="opt")
```

Die Lösungen dazu lauten:

```
> # allocate the sample size across strata: proportional allocation
> stratasamp(n=n_prop_int, Nh=c(100,700,200), Sh=c(6,30,12), type="prop")

Stratum  1   2  3
Size    39 273 78
>
> # allocate the sample size across strata: optimal allocation
> stratasamp(n=n_opt_int, Nh=c(100,700,200), Sh=c(6,30,12), type="opt")

Stratum 1   2  3
Size    8 297 34
```

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Selbstbeurteilung (Multiple-Choice-Fragen und Antworten)	Erhebungen auf der Grundlage von Stichproben: a) sparen Ressourcen im Vergleich zu einer zählungsbasierten Erhebung b) ermöglichen eine umfassende Untersuchung einer Population c) Beide Antworten sind richtig Der Stichprobenumfang wird beeinflusst durch: a) Die Fehlermarge und das Konfidenzintervall b) Das angewandte Stichprobenverfahren c) Beide Antworten sind richtig Bei der proportionalen Zuteilung wird der Stichprobenumfang auf die Schichten basierend auf folgendem Faktor verteilt: a) Die Varianz in jeder Schicht b) Die Größe der einzelnen Schichten c) Der Mittelwert für jede Schicht
Ressourcen (Videos, Verweislink)	
Verwandtes Material	
Verwandte PPT	
Literaturverzeichnis	NEWBOLD, P. et al. (2008): Statistik für Management und Wirtschaft, (6. Auflage) Ed. Prentice Hall. Kapitel 20, S. 763-784.
Zur Verfügung gestellt von	[Uniovi]

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.