Eine Datenbank für den Mathe-Unterricht

Dr. Hubert Langlotz

Dr. Wilfried Zappe

1

[1] https://st6.cannypic.com/thumbs/21/211995_352_canny_pic.jpg

Gliederung:

Übersicht über begleitende Dateien für den TI-Nspire CX CAS

Daten_Neugeborene_ti_tns

Daten_Neugeborene_rH_Geschlecht.tns

Daten_Neugeborene_Prognose_Geschlecht.tns

Daten_Neugeborene_Konfidenz_einfach.tns

Daten_Neugeborene_Konfidenz_Erwartungswert.tns

Daten_Neugeborene_Konfidenz_grafisch_mit prog.tns

Daten_Neugeborene_Konfidenz_grafisch-ohne prog.tns

Eine Datenbank für den Mathe-Unterricht

Vorbemerkungen

Emil Julius erblickte am 21. Juni 2020 im Suhler SRH-Klinikum das Licht der Welt. Bei seiner Geburt war er 47 Zentimeter groß und brachte 2490 Gramm auf die Waage.

Texas Instruments stellt hiermit eine Datenbank zur Verfügung, die interessierte Kolleginnen und Kollegen für ihren Mathe-Unterricht unentgeltlich nutzen können. Die Datenbank enthält die Angaben über Körpergröße, Körpergewicht und Geschlecht von über 2000 Neugeborenen. Über sechs Jahre lang wurden dazu Anzeigen aus Tageszeitungen ausgewertet. Die Anzeigen erscheinen etwa in wöchentlichem Rhythmus. Sicher kann man davon ausgehen, dass die Daten nur mit dem Einverständnis der Eltern veröffentlicht wurden. Sie enthalten mitunter auch Angaben über das Geburtsdatum und die Namen der Eltern. Aus Datenschutzgründen werden solche Angaben aber in der hier zur Verfügung gestellten Datenbank nicht weitergegeben. Ursprünglich erfasst wurden auch die Vornamen der Kinder. Sie wurden ebenfalls aus Datenschutzgründen entfernt. Vor der Entfernung der Vornamen wurde ermittelt, aus wie vielen Buchstaben und Zeichen der jeweilige Vorname zusammengesetzt ist. Ebenso wurden die Daten von Mehrlingsgeburten entfernt, um eine Unabhängigkeit zwischen den erhobenen Daten abzusichern.

Sie sehen im Folgenden einen Ausschnitt aus der Tabelle mit den Daten: [2]

	A nr	B größe	C masse	D geschlecht	E anzbu	F
=	=seq(k,k,1					
1	1	50	3.29	0	11	
2	2	52	3.73	1	6	
3	3	57	4.295	0	4	
4	4	50	2.8	1	10	

nr: Jeder Datensatz ist mit einer Nummer versehen.

größe: Enthält die Körpergröße in cm.

masse: Zeigt das Körpergewicht in kg an.

geschlecht: Eine „0“ steht für einen neugeborenen Jungen, eine „1“ für ein Mädchen.

anzbu: Jede Zahl gibt an, aus wie vielen Buchstaben und Zeichen der Vorname des Kindes zusammengesetzt ist.

Zusammengetragen wurden im Zeitraum von 2014 bis 2020 die Daten von Dr. Wilfried Zappe (Ilmenau). Sie beziehen sich in der Mehrheit auf Neugeborene aus dem Ilm-Kreis.

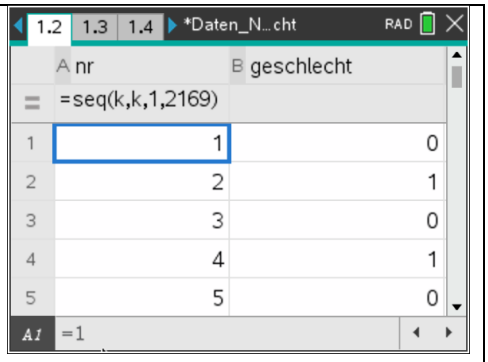
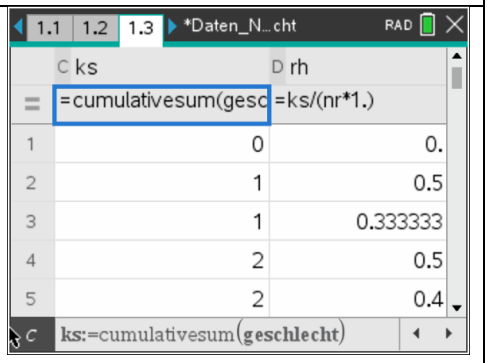
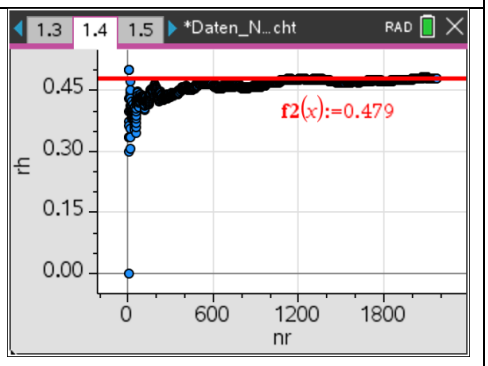
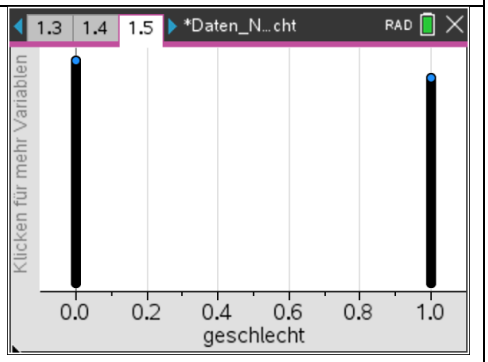
Die Kollegen Dr. Hubert Langlotz und Dr. Wilfried Zappe geben nun einige Anregungen dafür, was mit diesen Daten im Unterricht machbar ist. An dem Teil zur Statistik hat Kollege Tobias Kellner mitgearbeitet. Zur Programmierung bei Konfidenzintervallen konnten wir auf die Unterstützung des Kollegen Sebastian Rauh bauen, der außerdem den Text nochmal kritisch durchgesehen hat und schöne Ideen zum Thema beigesteuert hat. Zu besonderem Dank sind wir dem Kollegen Dr. Andreas Prömmel verpflichtet, der sehr sorgfältig alles vorhandene Material gesichtet und geprüft hat.

Hinweise oder Ergänzungen zu unseren Ausführungen nehmen wir gern entgegen. (info@t3deutschland.de)

[2] Quelle: „Freies Wort“ vom 04.07.2020

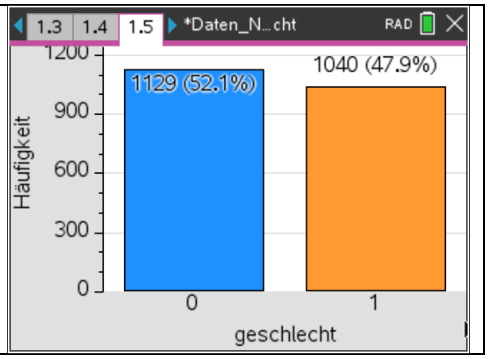
Stabilisierung von relativen Häufigkeiten und Durchschnittswerten

Mit zunehmender Anzahl von Einzeldaten ist eine Stabilisierung der relativen Häufigkeiten der Mädchen- bzw. Jungengeburten sowie der durchschnittlichen Größe und Masse neugeborener Kinder zu beobachten. Diese Prozesse können veranschaulicht werden. Sie tragen z. B. zu einem tieferen Verständnis des Wahrscheinlichkeitsbegriffes bei.

Kopieren Sie die Datei „Daten Neugeborene_2“ und fügen Sie diese Datei in die Anwendung „Lists&Spreadsheet“ ein. Löschen Sie die Spalten „größe“ und „masse“, sodass zunächst nur die Spalten A: „nr“ und B: „geschlecht“ verbleiben.	
Bezeichnen Sie die Spalte C mit dem Namen „ks“ für kumulierte Summe und geben Sie in die Zelle darunter den Befehl „=cumsum(geschlecht)“ ein. Der Rechner ergänzt diese Anweisung automatisch mit dem vollständigen Befehl „cumulativesum(geschlecht)“ und berechnet diese kumulierten Summen. In der Spalte D tragen Sie den Namen „rh“ für relative Häufigkeit ein und fügen in die Zelle darunter die Anweisung „=ks/(nr*1.)“ ein, um die relativen Häufigkeiten zu berechnen. Der Faktor „1.“ wird verwendet, damit die relativen Häufigkeiten als Dezimalbrüche angezeigt werden.	
Fügen Sie die Anwendung „Data&Statistics“ ein. Weisen Sie der horizontalen Achse die Variable „nr“ zu und der vertikalen Achse die Variable „hr“. Sie können die Stabilisierung der relativen Häufigkeiten gut erkennen: Je höher n, desto weniger schwanken die relativen Häufigkeiten um einen gewissen festen Wert, der hier bei ca. 0,479 liegt, d. h. hier, dass der Anteil der Mädchengeburten in dieser Stichprobe bei ca. 47,9% liegt.	
Eine rasche Übersicht über den Anteil der Mädchen- und Jungengeburten unter allen erfassten Neugeborenen erhalten Sie über die Darstellung der Liste „geschlecht“ in der Anwendung „Data&Statistics“.	

Die Anweisungen „Kategorisches X erzwingen“ und „Alle Bezeichnungen anzeigen“ ergeben den nebenstehenden Bildschirm.
Setzen Sie dazu den Kursor auf die Bezeichnung der horizontalen Achse und wählen Sie mit <ctrl><menü> die oben genannten Befehle.

Unter den 2169 erfassten Datensätzen sind 47,9% (1040) Mädchengeburten registriert.

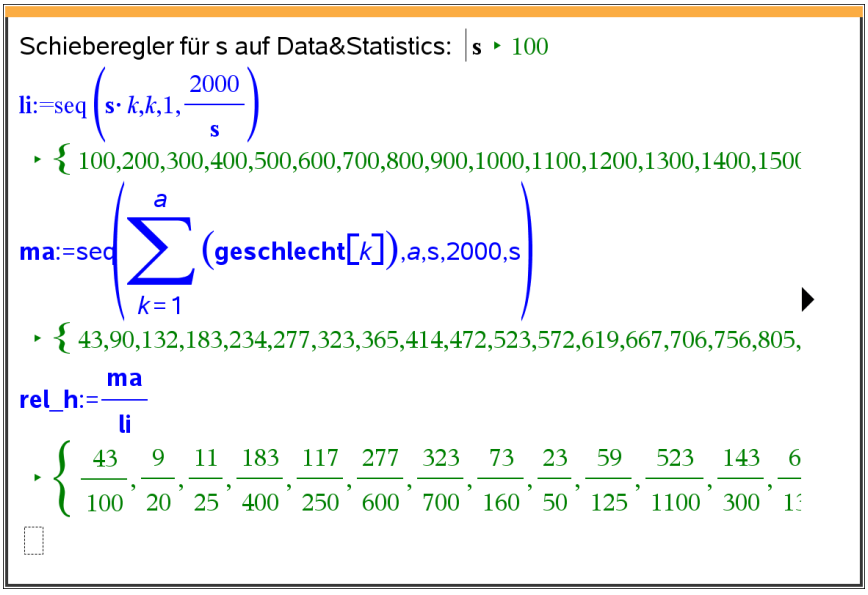


Aufgaben:

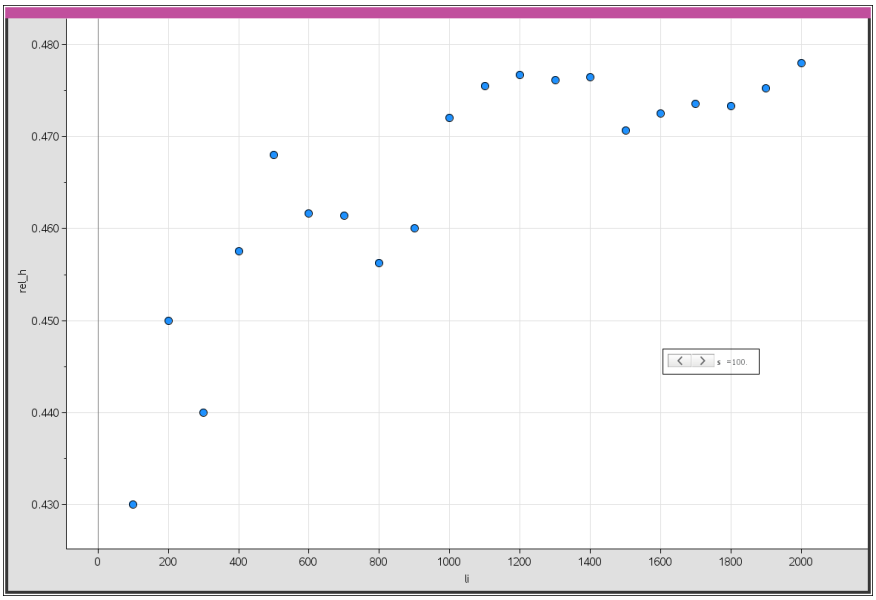
Untersuchen Sie auf analogem Wege die Stabilisierung der durchschnittlichen Körpergröße bzw. des durchschnittlichen Geburtsgewichts Neugeborener.

Ergänzungen:

Um die etwas unübersichtlich wirkende Darstellung von über 2000 einzelnen relativen Häufigkeiten zu verbessern, lassen sich die Stabilisierungseffekte auch durch die Zusammenfassung von zunächst den ersten 100, dann den ersten 200 zu Partialsummen aufeinanderfolgenden Einzelwerten usw. veranschaulichen. Jede dieser Partialsummen wird dann durch die Anzahl ihrer Summanden geteilt.



Die relativen Häufigkeiten sind hier nicht als Dezimalzahlen angegeben, weil sich durch die Darstellung in Brüchen ihre Herkunft besser erklären lässt.



Die grafische Darstellung muss ggf. durch *Zoom-Daten* angepasst werden.

Statistischen Auswertungen von Daten

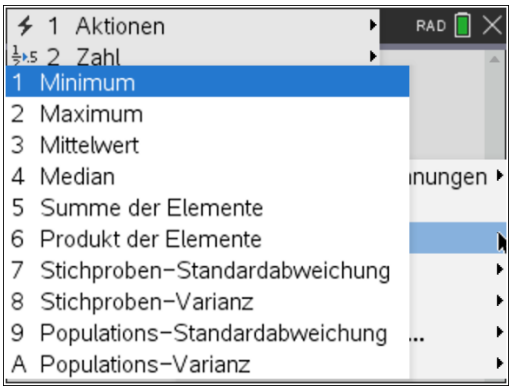
Das Merkmal „Geschlecht“ liegt in zwei Ausprägungen vor, die als „Klassenwerte“ bezeichnet werden:

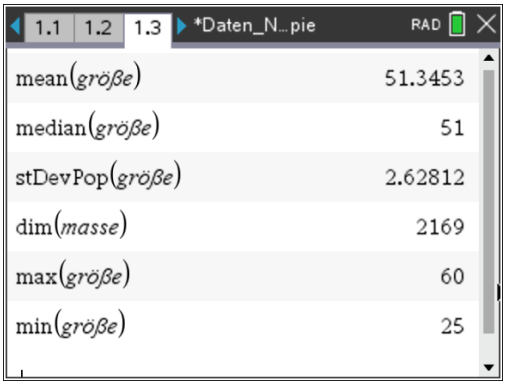
Die Klassenwerte müssen als Zahlen eingegeben werden, um eine nach Klassen getrennte Auswertung vornehmen zu können.

Geschlecht	Klassenwert
männlich	0
weiblich	1

Ermittlung von statistischen Kenngrößen über Listenbefehle

Von großem Interesse für die Bewertung von Daten sind ihre Kenngrößen, wie das arithmetische Mittel, die Standardabweichung, der Median und andere. Der CAS-Rechner TI-Nspire bietet dazu verschiedene Möglichkeiten an. Ein einfacher Zugriff kann über die Anwendung *Calculator - Menü - Statistik – Listen Mathematik* erfolgen. Er bietet folgende Optionen, die eigentlich selbsterklärend sind:





Die Anweisung „dim“ gibt die Anzahl der Listenelemente zurück.

Statistik mit einer Variablen für alle Neugeborenen

Komfortabler gelingt die Anzeige von Kenngrößen mit der Anwendung *Calculator - Menü – Statistik – Statistische Berechnungen – Statistik mit einer Variablen*.

Es öffnet sich ein Fenster, das die Anzahl der Listen abfragt. Wir lassen es zunächst bei der voreingestellten „1“ und drücken „ok“. Nun öffnet sich ein weiteres Fenster (siehe Screenshot). Um die statistischen Werte für alle Neugeborenen bezüglich des Merkmals „Größe“ zu erhalten, wählen wir als x1-Liste die Variable „größe“ aus, ändern weiter nichts an den Voreinstellungen und drücken „ok“. In den Spalten E und F werden Kenngrößen und ihre aktuellen Werte angezeigt. Die farbige Unterlegung wurde nachträglich vorgenommen.

	A nr	B größe	C masse	D geschlecht	E	F
=	=seq(k,k,1,2169)					=OneVar('größe,1): CopyVar Stat
1	1	50	3.29	0	Titel	Statistik mit einer Variable
2	2	52	3.73	1	$\bar{x}$	51.3453
3	3	57	4.295	0	Σx	111368.
4	4	50	2.8	1	Σx^2	5.73321E6
5	5	57	4.35	0	$sx := s_{n-1}x$	2.62873
6	6	52	3.69	1	$\sigma x := \sigma_n x$	2.62812
7	7	32	3.25	0	n	2169.
8	8	52	3.612	0	MinX	25.
9	9	48	2.275	0	Q_1X	50.
10	10	50	3.2	0	MedianX	51.
11	11	51	3.15	1	Q_3X	53.
12	12	53	3.06	0	MaxX	60.
13	13	50	2.96	0	$SSX := \Sigma(x-\bar{x})^2$	14981.4

Bedeutung der angezeigten Statistikvariablen

	$\bar{x}$ arithmetisches Mittel
$\bar{x}$	$\sum x$ Summe der Werte
Σx	$\sum x^2$ Summe der quadrierten Werte
Σx^2	sx Standardabweichung der Stichprobe (korrigiert)
$sx := s_{n-1}x$	σ_x Standardabweichung der Grundgesamtheit (unkorrigierte)
$\sigma x := \sigma_n x$	n Anzahl der Werte
n	kleinster Wert
MinX	erstes Quartil
Q_1X	Median der Werte
MedianX	drittes Quartil
Q_3X	größter Wert
MaxX	Summe der quadrierten Abweichungen der Einzelwerte vom arithmetischen Mittel
$SSX := \Sigma(x-\bar{x})^2$	

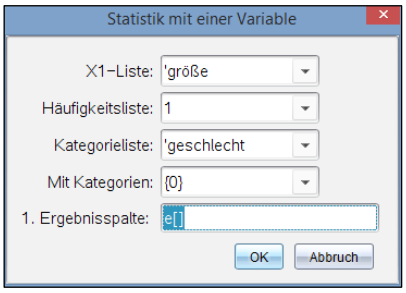
Aufgabe:

Bestimmen Sie auf analogem Wege die Kenngrößen der Listen „masse“ und „geschlecht“.

Statistik mit einer Variablen getrennt nach Kategorien, also wenn z. B. aus der Gesamtheit der Daten „größe“ nur die statistischen Angaben für die neugeborenen Jungen berechnet werden sollen.

Menü – Statistik – Statistische Berechnungen – Statistik mit einer Variablen wählen.

Es öffnet sich ein Fenster, das die Anzahl der Listen abfragt. Wir lassen es wieder bei der voreingestellten „1“ und drücken „ok“. Nun öffnet sich ein weiteres Fenster (siehe Bildschirmabdruck). Um die statistischen Werte für alle neugeborenen Jungen bezüglich des Merkmals „Größe“ zu erhalten, wählen wir als x1-Liste die Variable „größe“ aus. Die Häufigkeitsliste bleibt bei der voreingestellten „1“. Als Kategorieliste wählen wir die Variable „geschlecht“, tragen in das Feld Kategorien {0} für „männlich“ ein. Als Ergebnisspalte kann man eine der noch freien Spalten wählen. Wir nehmen hier wieder die Spalte E und lassen die vorigen Ergebnisse durch Drücken von „ok“ überschreiben.



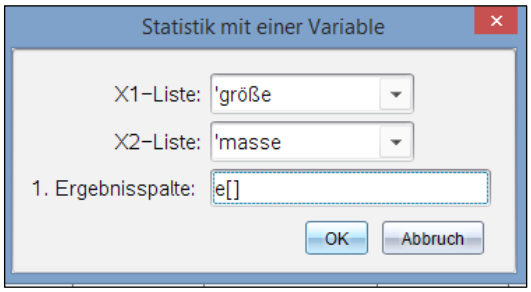
	A nr	B größe	C masse	D geschlecht	E	F
=	=seq(k,k,1,2169)					=OneVar('größe,1,'geschlecht,0):
1	1	50	3.29	0	Titel	Statistik mit einer Variable
2	2	52	3.73	1	$\bar{x}$	51.6594
3	3	57	4.295	0	Σx	58323.5
4	4	50	2.8	1	Σx^2	3.02138E6
5	5	57	4.35	0	$sx := s_{n-1}x$	2.73169
6	6	52	3.69	1	$\sigma x := \sigma_n x$	2.73048
7	7	32	3.25	0	n	1129.
8	8	52	3.612	0	MinX	25.
9	9	48	2.275	0	Q_1X	50.
10	10	50	3.2	0	MedianX	52.
11	11	51	3.15	1	Q_3X	53.
12	12	53	3.06	0	MaxX	60.
13	13	50	2.96	0	$SSX := \Sigma(x-\bar{x})^2$	8417.3

Aufgabe:

Ermitteln Sie analog die Statistikvariablen für die neugeborenen Mädchen.

Statistik mit einer Variablen für mehrere Listen

Menü – Statistik – Statistische Berechnungen – Statistik mit einer Variablen wählen.



Es öffnet sich ein Fenster, das die Anzahl der Listen abfragt. Wir wollen die Statistikvariablen für die Größe und die Masse gleichzeitig bestimmen lassen. Die voreingestellten „1“ wird durch eine „2“ ersetzt und „ok“ gedrückt. Nun öffnet sich ein weiteres Fenster (siehe Bildschirmabdruck). Um die statistischen Werte für alle Neugeborenen bezüglich der Merkmale „Größe“ und „Masse“ zu erhalten, wählen wir als x1-Liste die Variable „größe“ und als x2-Liste „masse“ aus. Für die 1. Ergebnisspalte wird Spalte E gewählt und „ok“ gedrückt.

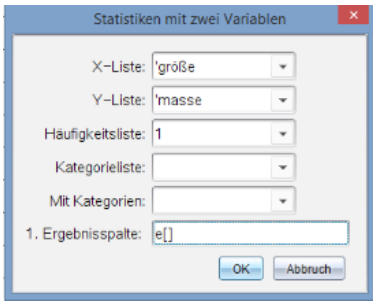
Die Statistikvariablen für jede Liste werden in getrennten Spalten angezeigt.

	B größe	C masse	D geschlecht	E	F	G
=	ɟ(k,k,1,2169)				=OneVar('größe,1): CopyVar Stat	=OneVar('masse,1): CopyV
1	1	50	3.29	0	Titel	Statistik mit einer Variable	Statistik mit einer Variable...
2	2	52	3.73	1	$\bar{x}$	51.3453	3.40803
3	3	57	4.295	0	Σx	111368.	7392.01
4	4	50	2.8	1	Σx^2	5.73321E6	25719.4
5	5	57	4.35	0	$sx := s_{n-1}x$	2.62873	0.493142
6	6	52	3.69	1	$\sigma x := \sigma_n x$	2.62812	0.493029
7	7	32	3.25	0	n	2169.	2169.
8	8	52	3.612	0	MinX	25.	0.45
9	9	48	2.275	0	Q_1X	50.	3.09
10	10	50	3.2	0	MedianX	51.	3.41
11	11	51	3.15	1	Q_3X	53.	3.72
12	12	53	3.06	0	MaxX	60.	5.935
13	13	50	2.96	0	$SSX := \Sigma(x-\bar{x})^2$	14981.4	527.235

Für eine Auswertung nach dem Klassenwert gibt es hier keine Abfrage, dies scheint für eine Anzahl von mehr als einer Liste nicht möglich zu sein.

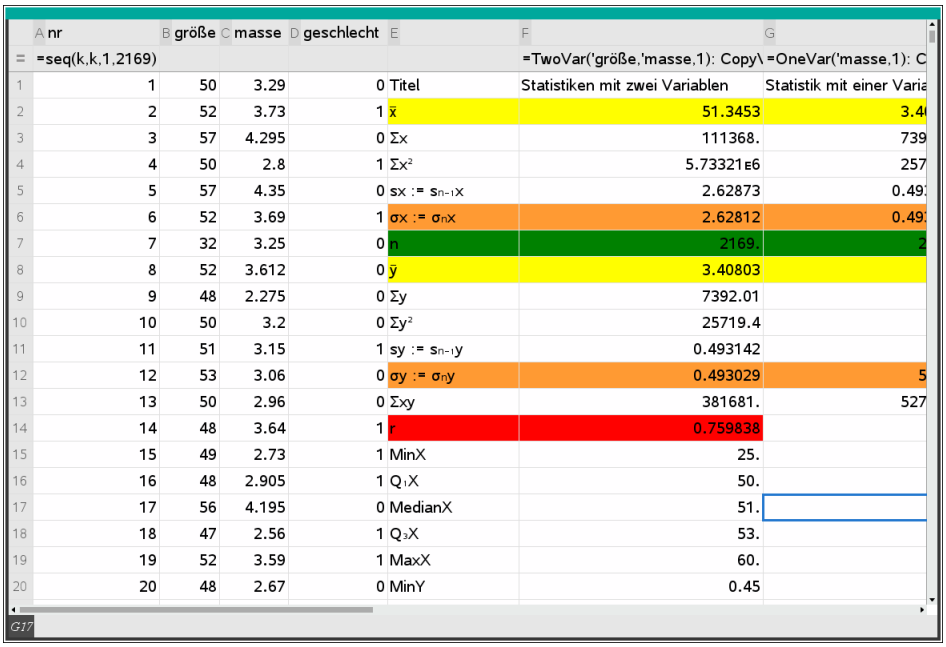
Statistik mit zwei Variablen; gibt es einen Zusammenhang zwischen Größe und Geburtsgewicht?

Menü – Statistik – Statistische Berechnungen – Statistik mit zwei Variablen wählen.



Es öffnet sich ein Fenster, das die Anzahl der Listen abfragt. Wir wollen die Statistikvariablen für die Größe und die Masse sowie ihren Zusammenhang bestimmen lassen. Die voreingestellte „1“ wird belassen. Die Auswertung soll nur für alle erfassten Neugeborenen erfolgen, deshalb bleiben die Felder „Kategorieliste“ und „Mit Kategorien“ leer. Für die 1. Ergebnisspalte wird Spalte E gewählt und „ok“ gedrückt. (Eine Auswertung nach Kategorien wäre aber auch möglich.)

Die Statistikvariablen für jedes Merkmal werden in ein und derselben Spalte angezeigt. Es sind so viele Variablen, dass man den Bildschirm nach unten scrollen muss, um alle zu sehen.



	A nr	B größe	C masse	D geschlecht	E	F	G
=	=seq(k,k,1,2169)					=TwoVar('größe,'masse,1): Copy\	=OneVar('masse,1): C
1	1	50	3.29	0	Titel	Statistiken mit zwei Variablen	Statistik mit einer Varia
2	2	52	3.73	1	$\bar{x}$	51.3453	3.4
3	3	57	4.295	0	Σx	111368.	739
4	4	50	2.8	1	Σx^2	5.73321E6	257
5	5	57	4.35	0	$sx := s_{n-1}x$	2.62873	0.49
6	6	52	3.69	1	$\sigma x := \sigma_n x$	2.62812	0.49
7	7	32	3.25	0	n	2169.	2
8	8	52	3.612	0	$\bar{y}$	3.40803	
9	9	48	2.275	0	Σy	7392.01	
10	10	50	3.2	0	Σy^2	25719.4	
11	11	51	3.15	1	$sy := s_{n-1}y$	0.493142	
12	12	53	3.06	0	$\sigma y := \sigma_n y$	0.493029	5
13	13	50	2.96	0	Σxy	381681.	527
14	14	48	3.64	1	r	0.759838	
15	15	49	2.73	1	MinX	25.	
16	16	48	2.905	1	Q_1X	50.	
17	17	56	4.195	0	MedianX	51.	
18	18	47	2.56	1	Q_3X	53.	
19	19	52	3.59	1	MaxX	60.	
20	20	48	2.67	0	MinY	0.45	

G17

Die Statistikvariable „r“ ist der „Korrelationskoeffizient“. Er ist definiert durch[3]

$$\mathrm{Kor}_e(x,y) := \varrho_e(x,y) := r_{xy} := \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

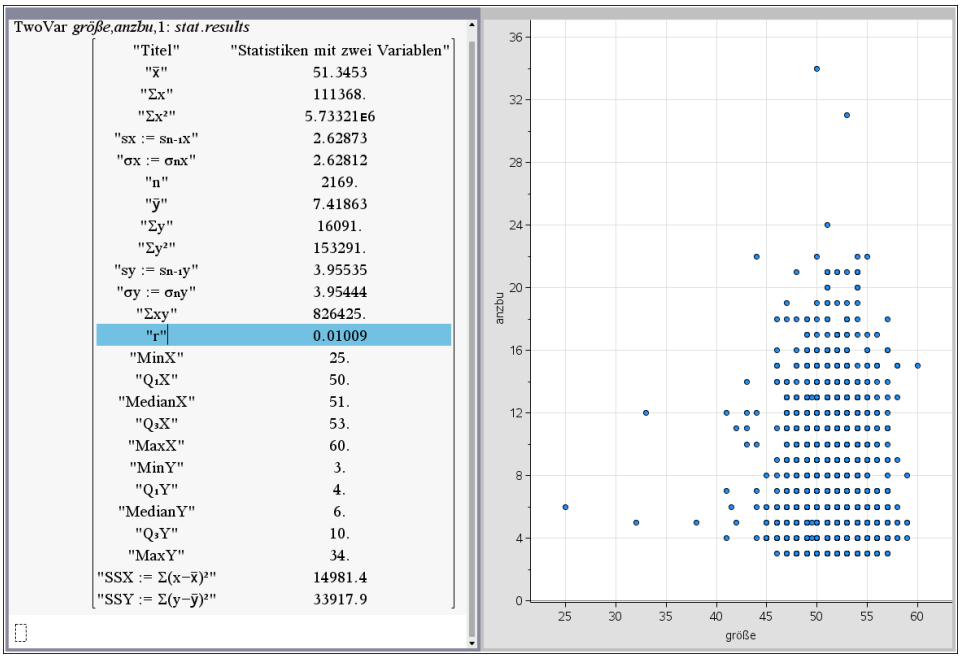
Er wird verwendet, um festzustellen, wie hoch ein Zusammenhang zwischen zwei numerischen oder quantitativen Merkmalen ist. Er bezeichnet das Maß für die Richtung und Stärke einer statistischen Beziehung zwischen zwei Merkmalen.
Der Korrelationskoeffizient nimmt Werte zwischen – 1 und +1 an. Ein Wert von +1 bedeutet, dass bei steigenden Werten des ersten Merkmals die des zweiten Merkmals maximal steigen, und umgekehrt. Ein Wert von -1 bedeutet, dass bei steigenden Werten des ersten Merkmals die Werte des anderen maximal sinken.
Man kann sogar zeigen, dass genau dann, wenn $r = \pm 1$ ein fast sicherer affin linearer Zusammenhang zwischen x und y besteht.
Ein Wert von r = +0,6 bedeutet, dass ein mittlerer positiver Zusammenhang besteht, ein Wert von +0,2, dass ein kleiner positiver Zusammenhang vorliegt. Hierbei ist zu erwähnen, dass der Koeffizient zwar etwas über die Korrelation aussagt, sich aus dem Ergebnis aber nicht der kausale Zusammenhang ableiten lässt. Als Beispiel ist hier die Besiedelung des österreichischen Südburgenlands durch Störche zu erwähnen. Diese korreliert zwar positiv mit der Geburtenrate, aber es lässt sich trotzdem kein ursächlicher Zusammenhang ableiten.

Der Korrelationskoeffizient r ist für Größe und Masse bei den hier untersuchten neugeborenen Kindern ca. 0,76. Beide Merkmale weisen also eine mittlere Korrelation auf. Man darf hier sicher auch einen kausalen Zusammenhang zwischen Größe und Körpergewicht vermuten.[4]

Auf analogem Wege lässt sich z. B. untersuchen, welche Korrelation zwischen der Größe und der Anzahl der Buchstaben des Vornamens besteht:

[3] http://de.wikipedia.org/wiki/Korrelationskoeffizient

[4] Vgl. z. B. Daten und Zufall im Mathematikunterricht, Cornelsen Verlag 2012, S. 95ff.



Es verwundert nicht, dass der Korrelationskoeffizient hier mit rund 0,01 nahe bei Null liegt.

Grafische Darstellung von Daten durch Boxplots

Durch Aufrufen der Anwendung *Data&Statistics* kann man die Daten auch in verschiedenen Varianten grafisch darstellen. Von beonderer Bedeutung sind dabei Boxplots.

Öffnen Sie Data&Statistics.

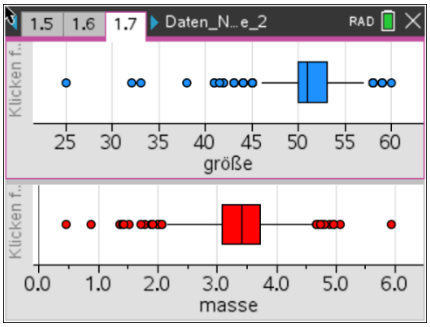
Drücken Sie <tab> und wählen Sie die Sie interessierende Liste aus.

Bestätigen Sie mit <enter>.

Drücken Sie <ctrl><menü> und wählen Sie *Box Plot*.

Sie erhalten die zugehörige *Kästchengrafik*.

Durch Überstreichen der Kästchengrafik mit dem Kursor werden das untere Quartil, der Median und das obere Quartil sowie die Werte der Ausreißer numerisch angezeigt.



Prognoseintervalle

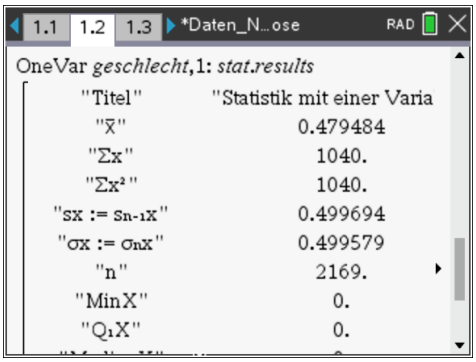
In den Jahren 2014 bis 2019 wurden in Deutschland 2 356 752 Jungen und 2 238 378 Mädchen geboren.

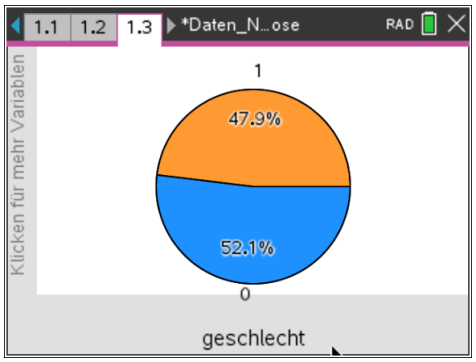
Jahr	Jungen	Mädchen
2019	399 292	378 798
2018	404 052	383 471
2017	402 510	382 374
2016	405 585	386 546
2015	378 478	359 097
2014	366 835	348 092
Summe	2 356 752	2 238 378

Quelle: https://de.statista.com/statistik/daten/studie/880778/umfrage/anzahl-der-geburten-in-deutschland-nach-geschlecht/

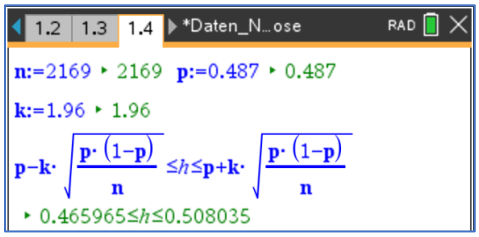
Der Anteil der Mädchengeburten in der Gesamtheit in Deutschland betrug danach in diesem Zeitraum $\frac{2\,238\,378}{2\,356\,752+2\,238\,378} \approx 0{,}487$.

Der Anteil der Mädchengeburten in der in den Jahren 2014 bis Mitte 2020 erhobenen Stichprobe (n = 2169, Liste „geschlecht“) beträgt 0,479.





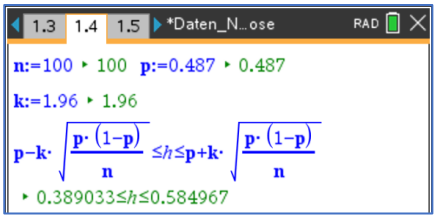
Wir prüfen, ob diese relative Häufigkeit im 95%-Prognoseintervall zu p = 0,487 liegt:



Das 95%-Prognoseintervall ist $0{,}466 \leq h \leq 0{,}508$. Die relative Häufigkeit h = 0,479 des Stichprobenergebnisses liegt in diesem Intervall. Das Stichprobenergebnis ist also statistisch verträglich mit p = 0,487.

Wir testen, wie viele von 20 (100) zufälligen Stichproben vom Umfang n = 100 aus der Liste „geschlecht“ statistisch verträglich mit dem 95%-Prognoseintervall sind:

Zunächst wird für n = 100 und p = 0,487 (Wahrscheinlichkeit für Mädchengeburten in Deutschland) das zugehörige 95%-Prognoseintervall $0{,}390 \leq h \leq 0{,}584$ bestimmt.



Genau eine zufällige Stichprobe vom Umfang n = 100 aus der Liste „geschlecht“ ohne Zurücklegen wird mit dem Befehl „randsamp (**geschlecht**, 100,1)“ bestimmt. Die relative Häufigkeit rh der Mädchengeburten in dieser Stichprobe kann durch die Summenbildung erfolgen, da die Mädchengeburten mit „1“ charakterisiert wurden. Im nebenstehenden Beispiel ergibt sich die Summe 52, sodass sich für die relative Häufigkeit rh feststellen lässt, dass sie im Prognoseintervall $0{,}390 \leq rh \leq 0{,}584$ liegt.

Eine Liste von 20 solcher Stichproben kann mit dem Befehl „seq“ erzeugt werden. Der Befehl „countif“ dient zum Abzählen derjenigen Listenelemente, die im 95%-Prognoseintervall liegen. Im nebenstehenden Beispiel wären das 19 von 20, also gerade 95% der Stichproben, die im 95%-Prognoseintervall liegen. Wird diese Berechnung wiederholt, können sich auch andere Werte ergeben, weil es sich ja um Zufallsversuche handelt. Aber eine Mittelwertbildung der Ergebnisse über eine größere Anzahl wird in der Nähe der 95%-Marke liegen. Um eine solche Wiederholung der Berechnung zu realisieren, genügt es,

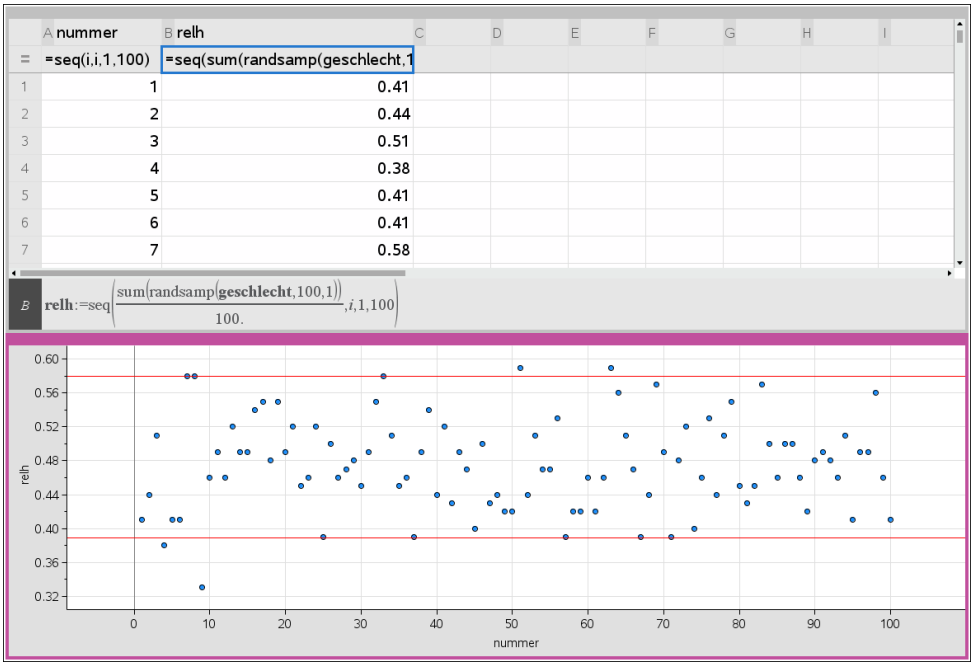
falls die Rechenschritte in der Anwendung „Notes“ formuliert wurden, den Kursor in die Anweisung „liste“ zu setzen und <enter> zu drücken.

Eine grafische Veranschaulichung für 100 solcher Stichproben kann über das nachfolgend skizzierte Vorgehen in der Anwendung „Lists&Spreadsheet“ realisiert werden. In der Spalte A werden die natürlichen Zahlen von 1 bis 100 erzeugt. Diese Liste erhält den Namen „nummer“. In der Spalte B werden 100 zufällige Stichproben ohne Wiederholung aus der Liste „geschlecht“ erzeugt und die relativen Häufigkeiten der Mädchengeburten in jeder Stichprobe berechnet. Diese Liste erhält den Namen „relh“. Die notwendigen Befehle lassen sich dem Screenshot entnehmen.

In der Anwendung „Data&Statistics“ wird jedem Element der Liste „nummer“ die zugehörige relative Häufigkeit aus der Liste „relh“ zugeordnet und das Wertepaar als Punkt dargestellt.

Zeichnet man noch die Grenzen des Prognoseintervalls als Graphen der konstanten Funktion y = 0,390 und y = 0,584 ein, lässt sich mit einem Blick erfassen, wie viele der Punkte außerhalb des Prognoseintervalls liegen.

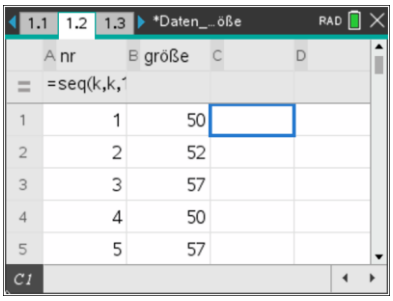
Auch hier lassen sich in der Tabellenkalkulation mit <ctrl><R> beliebig viele Wiederholungen erzeugen.

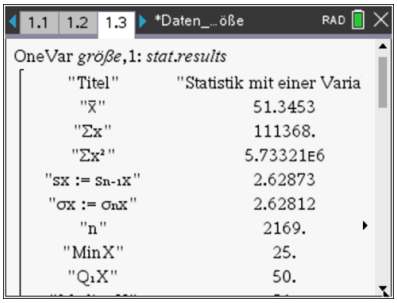


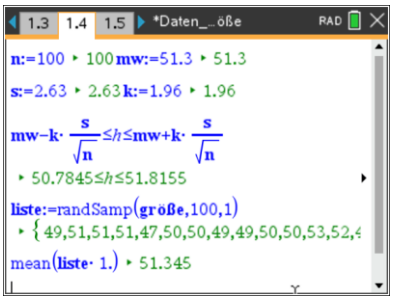
Prognoseintervalle für den Mittelwert einer normalverteilten Zufallsgröße:

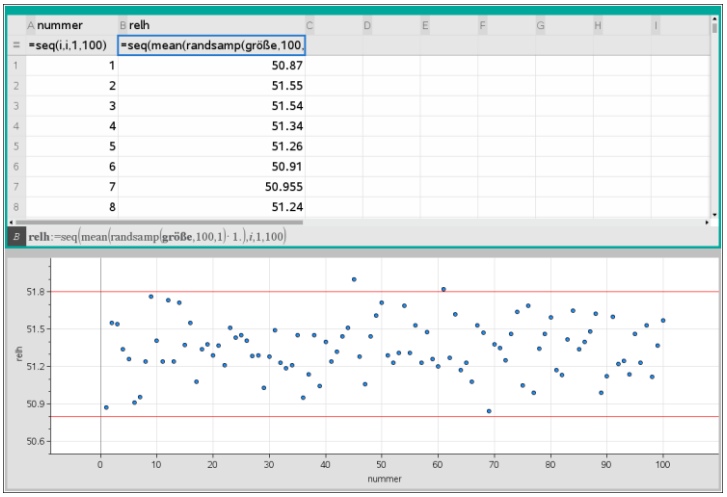
Für die annähernd normalverteilten Zufallsgrößen „masse“ und „größe“ lassen sich auf analogem Wege Prognoseintervalle für Mittelwerte bestimmen. Dazu werden der Mittelwert $\bar{x}$ und die Standardabweichung σ_x von der Gesamtheit aller erfassten Daten verwendet. Das Prognoseintervall hat dann die Gestalt $\bar{x} - k \cdot \frac{\sigma_x}{\sqrt{n}} \leq h \leq \bar{x} + k \cdot \frac{\sigma_x}{\sqrt{n}}$ (vgl. Bigalke/ Köhler „Mathematik, Gymnasiale Oberstufe, Qualifikationsphase Leistungskurs Q3“,Cornelsen, 2018, Seite 268).

Die folgenden Bildschirme veranschaulichen das Vorgehen:



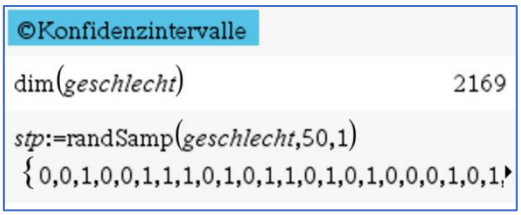






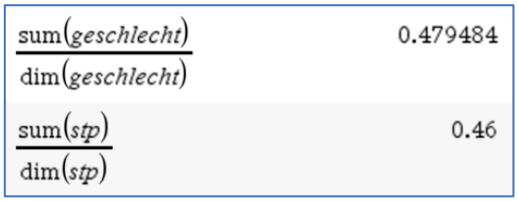
Konfidenzintervalle

Betrachtet man den gesamten Datensatz als Grundgesamtheit (hier n = 2169), so kann man auf verschiedenen Wegen Sicherheitsintervalle zu einer gegebenen Stichprobe (hier Stichprobengröße 50) ermitteln.



Im Screenshot erhält man als Punktschätzung für den Anteil der Mädchen in der Grundgesamtheit z. B. 0,46.

Der Anteil der Mädchen in der Grundgesamtheit beträgt ca. 0,48.



Ermittlung des 95% Konfidenzintervalls zur gegebenen Stichprobe auf verschiedenen Wegen.

Bei einer Bernoullikette der Länge n (wir können aufgrund der Größe der gewählten Stichproben - kleiner als 1/20 der Grundgesamtheit- dies annehmen) und der Trefferwahrscheinlichkeit p gilt nach der 1,96σ-Regel für die relative Haufigkeit h :

$$p - 1{,}96 \cdot \sqrt{\frac{p(1-p)}{n}} \leq h \leq p + 1{,}96 \cdot \sqrt{\frac{p(1-p)}{n}}.$$

$$\text{solve}\left(p-1.96\cdot\sqrt{\frac{p\cdot(1-p)}{50}} \leq \frac{23}{50} \leq p+1.96\cdot\sqrt{\frac{p\cdot(1-p)}{50}}, p\right)$$

$$0.329694 \leq p \leq 0.596014$$

Bei bekannter Trefferwahrscheinlichkeit p liefert diese Ungleichung ein 95%-Prognoseintervall für die relative Trefferhäufigkeit h. Umgekehrt kann man mit Hilfe der relativen Häufigkeit in einer Stichprobe hiermit den Wert von p schätzen.

Mit einem CAS kann immer direkt mit dieser Formel gearbeitet werden, allerdings existieren Näherungsformeln, die meist ausreichend gute Werte liefern.

1. Herleitung und Nutzung des $\frac{1}{\sqrt{n}}$ – Gesetzes:

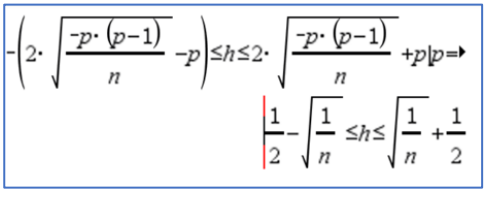
Für eine Sicherheitswahrscheinlichkeit von 95% kann

|h-p| mit dem $\frac{1}{\sqrt{n}}$ – Gesetz näherungsweise abgeschätzt werden durch $|h-p| \leq \frac{1}{\sqrt{n}}$.

Begründung:

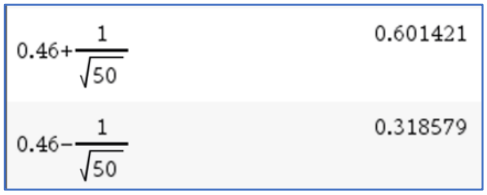
Die quadratische Funktion $f(p) = p \cdot (1-p)$ hat an der Stelle 0,5 ihren Scheitelpunkt und nimmt dort ihren größten Wert an.

Ersetzt man nun in der Doppelungleichung $1{,}96\sqrt{\frac{p\cdot(1-p)}{n}} \leq h \leq p + 1{,}96\sqrt{\frac{p\cdot(1-p)}{n}}$ die Variable p durch 0,5 und 1,96 durch 2, so erhält man das $\frac{1}{\sqrt{n}}$ – Gesetz.



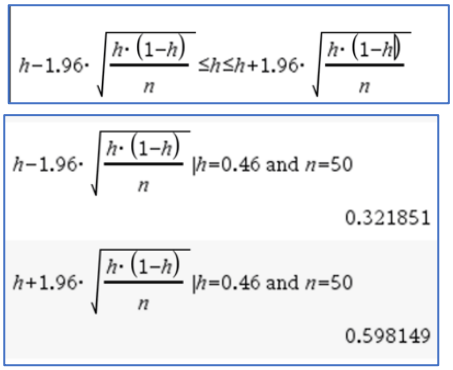
Die Anwendung dieser Näherungsformel liefert das Konfidenzintervall mit [0.32; 0.60]

Damit wird die „wirkliche“ Trefferwahrscheinlichkeit von 0,48 gut überdeckt.



2. Verwendung der Doppelungleichung

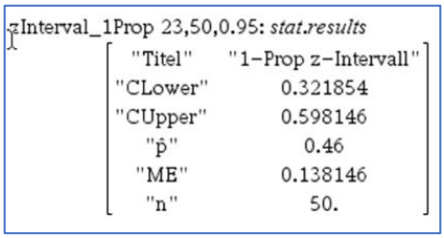
Auch hier bekommen wir ein Intervall, welches die „wirkliche Wahrscheinlichkeit“ überdeckt.



3. Nutzung der im CAS vorhandenen Näherungsformel

Die im Statistikmodul existierende Näherungsformel

liefert ebenfalls ein Sicherheitsintervall [0.33; 0.60].



Anmerkung:

Es wird hier das sogenannte Wald-Intervall berechnet, bei welchem zum einfacheren Berechnen p durch h abgeschätzt wird. Die Rundung erfolgt entsprechend der Ungleichheitszeichen in der Doppelungleichung.

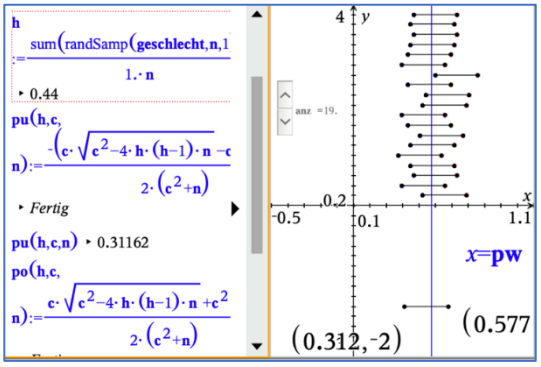
Erzeugung mehrerer Konfidenzintervalle unter Verwendung von „Notes“

Variante 1:

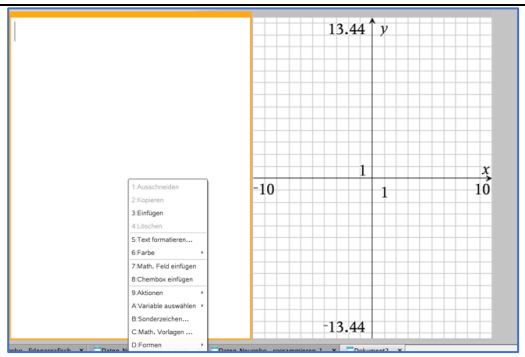
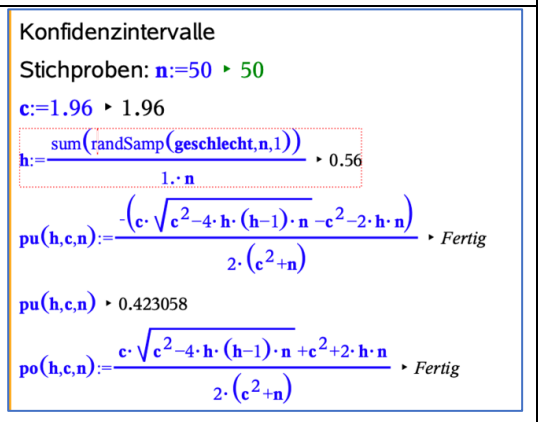
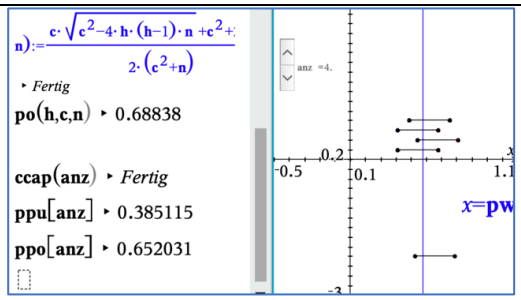
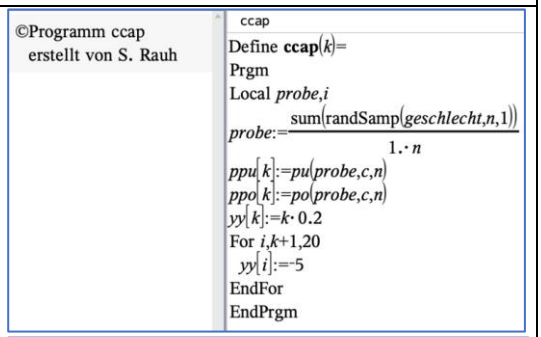
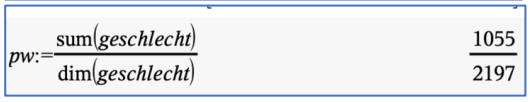
Mit Hilfe der Applikation „Notes“ lassen sich schnell viele verschiedene Konfidenzintervalle erzeugen.

Im linken Fenster sind die notwendigen Funktionen definiert und im rechten lassen sich mittels Schieberegler mehrere Konfidenzintervalle darstellen.

Hier wurden 20 Konfidenzintervalle erzeugt, von denen 19 die „gesuchte“ Wahrscheinlichkeit überdecken.



Anleitung zur Erzeugung der Darstellung

Zuerst erzeugt man sich ein zweigeteiltes Fenster, links fügt man eine Notes- und rechts eine Graphs-Applikation ein. Beachten Sie, dass man im Notes-Fenster normalen Text, Bilder und mathematische Formeln (Math-Feld-einfügen) einfügen kann.	
Die Stichprobengröße n ist ebenso wie der Faktor c (hier 1,96 für das 1,96-fache der Sigma-Umgebung) frei wählbar. Mit den Variablen pu und po werden die Grenzen des jeweiligen Konfidenzintervalls berechnet. (Man erhält die Terme durch Umformung der Doppelungleichung nach p.) Befindet sich der Cursor im Definitionsfeld der Variable h, so kann durch <Enter> immer eine neue Punktschätzung erzeugt werden.	
Möchte man mehrere Konfidenzintervalle gleichzeitig darstellen, so kann man dies z. B. mittels eines kleinen Programmes (ccap) unter Nutzung eines Schiebereglers (anz) erreichen.	
Das Programm ccap leistet Folgendes: Über den Parameter k wird die Nummer des zu erzeugenden Konfidenzintervalls übergeben, danach wird eine neue Stichprobe gezogen und die neue untere bzw. obere Grenze des Intervalls ermittelt (diese beiden Werte ppu[k] bzw. ppo[k] werden dann genutzt, um mit dem Wert yy[k] die Endpunkte der Strecken darzustellen, welche das Konfidenzintervall im Grafikfenster darstellen sollen. Die Strecken im Grafikfenster müssen dann einmal manuell erzeugt werden. Die For-Schleife sorgt nur dafür, dass immer nur die Intervalle von 1 bis k dargestellt werden, alle übrigen Strecken werden mittels des Befehls yy[i]=-5 in den nichtsichtbaren Bereich verschoben. Im Graphikfenster ist zusätzlich noch die „Wahrscheinlichkeit pw in der Grundgesamtheit“ dargestellt. Diese Variable muss vorher definiert werden.	 

Um auf die berechneten Werte zugreifen zu können, benötigt man noch ein Tabellenkalkulationsfenster, in welchem man in den drei dargestellten Spalten die durch das Programm erfassten Werte darstellt.	

	A ppu	B ppo	C yy	D
=				
1	=0.366442	0.633557	0.2	
2	0.293748	0.557668	-5	
3	0.329694	0.596014	-5	
4	0.403986	0.670306	-5	
5	0.347969	0.614885	-5	
6	0.31162	0.576942	-5	
7	0.31162	0.576942	-5	

Aufgaben:

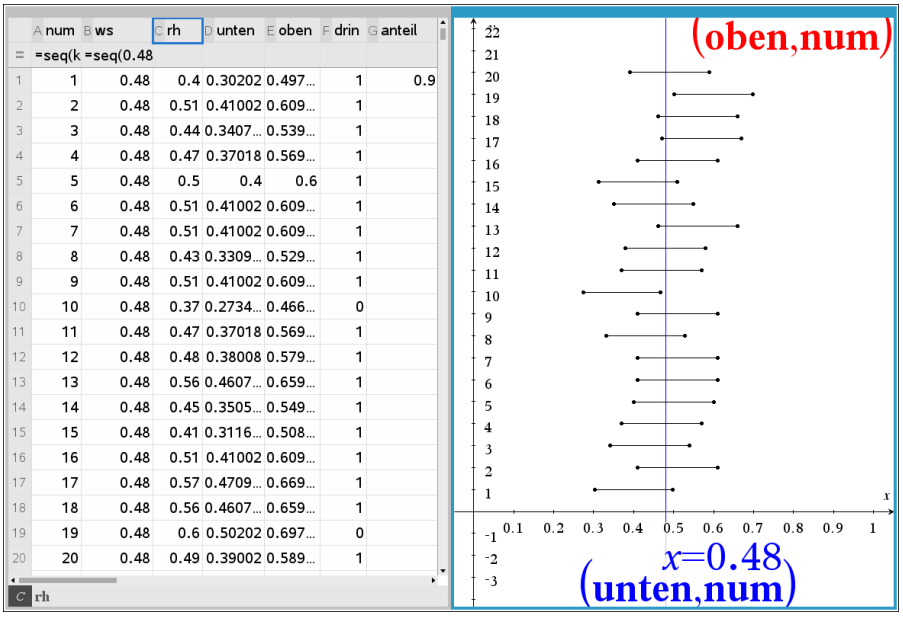
a) Testen Sie das Programm Daten_Neugeborene_Konfidenz_grafisch_mit prog.tns

b) Wählen Sie verschiedene Stichprobengrößen und untersuchen Sie den Zusammenhang zwischen Stichprobengröße und Intervalllänge.

Variante 2:

Einfaches Beispiel ohne Programmierung zur Veranschaulichung der Sicherheitswahrscheinlichkeit für Konfidenzintervalle

Es werden zwanzig 95,4%-Konfidenzintervalle für die Anteile der Mädchengeburten in Stichproben vom Umfang n = 50 aus der Liste „geschlecht“ erzeugt und grafisch veranschaulicht. Durch wiederholtes Durchführen lassen sich Mittelwerte für den Anteil der Konfidenzintervalle ermitteln, die die „unbekannte“ Wahrscheinlichkeit der Mädchengeburten in der Gesamtheit überdecken.



In diesem Beispiel erhält man zwei von zwanzig Konfidenzintervalle, die p = 0,48 nicht enthalten. 90% dieser Konfidenzintervalle überdecken also die Wahrscheinlichkeit p = 0,48.

Spalte	Bedeutung
A	Nummer der Simulation
B	angenommener Wert p = 0,48 der Wahrscheinlichkeit für eine Mädchengeburt
C	Anteil der Mädchengeburten in einer Stichprobe vom Umfang n = 50 aus der Liste *geschlecht*: $= sum(\frac{randsamp(geschlecht,50,1))}{50.}$

D	Näherungswert für die untere Grenze des Konfidenzintervalls: $= C1 - 2 \cdot \sqrt{\frac{C1 \cdot (1-C1)}{50}}$
E	Näherungswert für die obere Grenze des Konfidenzintervalls: $= C1 + 2 \cdot \sqrt{\frac{C1 \cdot (1-C1)}{50}}$
F	Legt den Wert 1 fest, wenn die binomialverteilte Zufallszahl im Konfidenzintervall liegt, sonst den Wert 0: $= when(D1 \leq B1 \leq E1,1,0)$
G	Gibt den Anteil der Konfidenzintervalle an, die mit p verträglich sind: $= \frac{sum(drin)}{20.}$

Hinweise:

Die Befehle für die Spalten C, D, E und F werden in Zeile 1 eingetragen und mit <Menü> <Daten> <Füllen> bis in die Zeile 20 als relative Zellbezüge kopiert. Die Spalten werden- wie oben zu sehen ist- bezeichnet. Die gewonnenen Daten werden als zwei Streudiagramme veranschaulicht. Die erhaltenen Punkte werden mit dem Geometriewerkzeug durch Strecken verbunden. Diese Strecken veranschaulichen die Konfidenzintervalle. Durch <CTRL> <R> in der Tabellenkalkulation können die Simulationen beliebig oft wiederholt werden. Zeichnet man noch die Gerade x = 0,48 ein, so lässt sich gut erkennen, welches der 20 Konfidenzintervalle p = 0,48 nicht überdeckt, falls ein solche Situation eintritt.

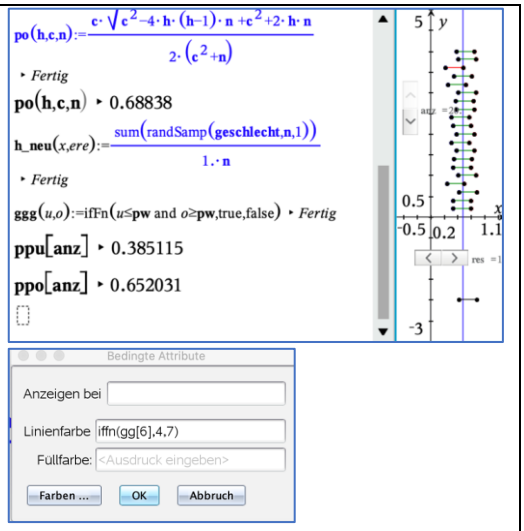
Aufgaben:

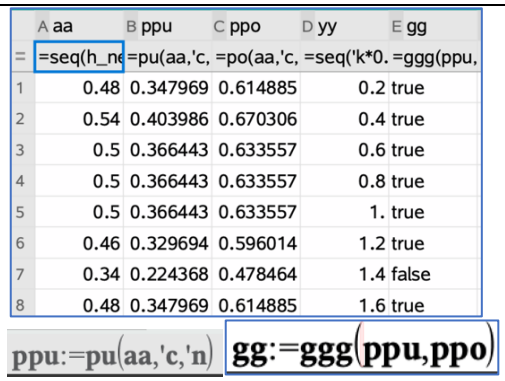
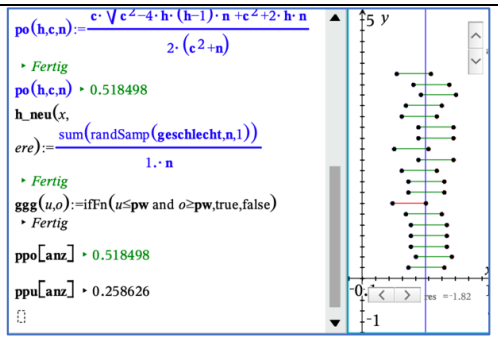
a) Realisieren Sie die Simulationen auf Ihrem CAS-Rechner.
b) Wiederholen Sie die Simulation mit <CTRL> <R> zehnmal.
c) Ermitteln Sie für Ihre zehn Wiederholungen einen Durchschnittswert für den Anteil der Konfidenzintervalle, die mit p = 0,48 statistisch verträglich sind.
d) Beschreiben Sie, wie die Simulation für andere Werte von p oder n angepasst werden kann.

Variante 3:

Konfidenzintervalle für Mädchengeburten ohne Programmierung

(Daten_Neugeborene_Konfidenz_grafisch-ohne prog.tns)

Um ohne Programmierung auszukommen, benötigt man z. B. noch einen weiteren Schieberegler (res), der für die Aktualisierung der Daten im Lists&Spreadsheetfenster zuständig ist. Auf diesen wird dann im L&S-Fenster über die Variable h_neu zugegriffen. Die Variable ggg dient nur dazu, jene Konfidenzintervalle rot zu färben, die die wirkliche Wahrscheinlichkeit nicht überdecken. (Hierzu müssen bei allen Strecken die Bedingungen zum Zeichnen eingetragen werden.	

Erläuterung zu den einzelnen Spalten: aa: Berechnung von 20 neuen Punktschätzungen ppu, ppo: Intervallgrenzen ermitteln yy: y-Koordinate zuordnen ggg: boolesche Variable, um die Strecken rot färben zu können, falls die Prüfung den Wert false ergibt Die notwendigen Formeln kann man der tns-Datei entnehmen.	
Ein Klick auf den Schieberegler „res“ erzeugt 20 neue Intervalle, welche man sich dann durch Klicken auf den Schieberegler „anz“ anzeigen lassen kann. Der Schieberegler „res“ wird benötigt, um in der Definition der Variable „aa“ eine Erneuerung zu erzeugen. Die notwendigen Formeln kann man der tns-Datei entnehmen.	

Konfidenzintervalle für den Erwartungswert einer Zufallsgröße:

Für die quantitativen Zufallsgrößen „masse“ und „größe“ lassen sich auf analogem Wege Konfidenzintervalle für Erwartungswerte bestimmen. Dazu werden der Mittelwert $\bar{x}$ und die Standardabweichung σ_x der in einer Stichprobe erfassten Daten verwendet.
Der Parameter k beschreibt das angestrebte Sicherheitsniveau, z. B. gilt k = 1,96 für ein Sicherheitsniveau von 95%.
Das Konfidenzintervall hat dann die Gestalt $\mu - k \cdot \frac{\sigma_x}{\sqrt{n}} \leq \bar{x} \leq \mu + k \cdot \frac{\sigma_x}{\sqrt{n}}$.
(vgl. Bigalke/ Köhler „Mathematik, Gymnasiale Oberstufe, Qualifikationsphase Leistungskurs Q3“,Cornelsen, 2018, Seite 268).

Der angegebenen Literatur entnimmt man noch folgenden Hinweis:

„Diese Näherungsverfahren darf nur angewendet werden, wenn der Stichprobenumfang n mindestens 30 beträgt und die Stichprobe maximal 5% der Grundgesamtheit umfasst.“ (ebenda, Fußnote auf Seite 268)

Betrachten wir den gegebenen Datensatz der Körpergröße Neugeborener als Stichprobe für alle Neugeborenen in Deutschland im gleichen Zeitraum, dann ergibt sich danach als Erwartungswert für die Körpergröße Neugeborener das 95%-Konfidenzintervall

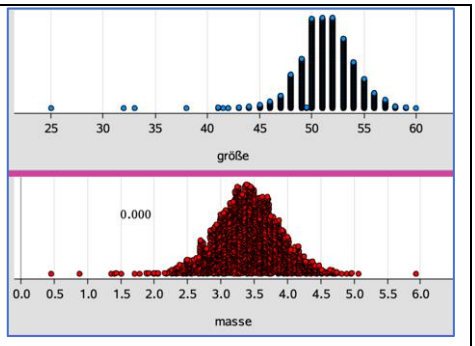
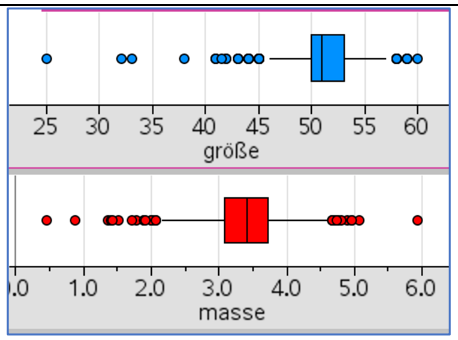
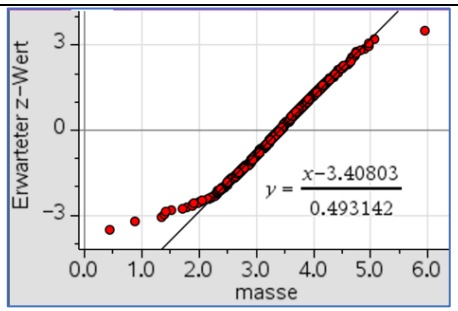
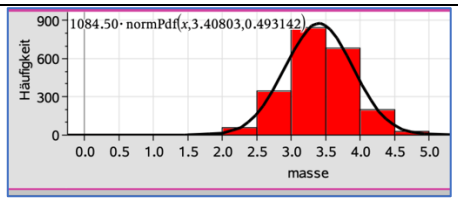
$51{,}24 \leq \mu \leq 51{,}45.$

Aufgabe:

Ermitteln Sie ein 95%-Konfidenzintervall für den Erwartungswert des Körpergewichtes Neugeborener.

Daten auf Normalverteilung prüfen

Mit schulischen Mitteln können wir nur qualitative Aussagen treffen, ob eine gegebene Verteilung annähernd normalverteilt ist oder nicht.

Darstellung als Punktdiagramm Zunächst kann man die Daten in einem Punktdiagramm darstellen, um eine erste Vorstellung der Verteilung zu bekommen. Eine glockenförmige Verteilung liefert ein erstes Indiz für eine mögliche Normalverteilung.	
Nutzung der Boxplots Durch Veränderung des Plottyps auf „Boxplot“ können beide Verteilungen neu dargestellt werden. Umso symmetrischer ein Boxplot ist, desto besser kann man die Verteilung durch die Normalverteilung nähern. Man sieht hier z. B., dass die Masse anscheinend besser durch eine Normalverteilung angepasst werden kann, als die Größe.	
Normal Wahrscheinlichkeitsdiagramm Mit ctrl menu wird der Befehl ***Normal Wahrscheinlichkeitsdiagramm*** ausgelöst. Die Datenpunkte werden um eine Gerade angeordnet. Je dichter sie an der Geraden liegen, desto besser lassen sich die Daten durch eine Normalverteilung modellieren. Der gleichzeitig angezeigten Geradengleichung (hier: $y = \frac{x-3{,}40803}{0{,}493142}$) lassen sich der Erwartungswert (3,41) und die Standardabweichung (0,49) entnehmen.	
Normalverteilung anzeigen Zunächst ändert man den Plottyp auf„Histogramm“. Mit *Menü – Analysieren – Normal Pdf anzeigen* kann eine Glockenkurve mit zugehöriger Gleichung erzeugt werden. Die Werte für Erwartungswert und Standardabweichung stimmen mit den vorher erzeugten überein.	
Sigma-Regeln nutzen Gemäß der Sigma-Regeln müssen in der Ein-Sigma-Umgebung des Erwartungswertes ca. 68% der Werte liegen, für die Aufgabe wäre dies das Intervall [3,41-0,49; 3,41+0,49] = [2,92; 3,9].