

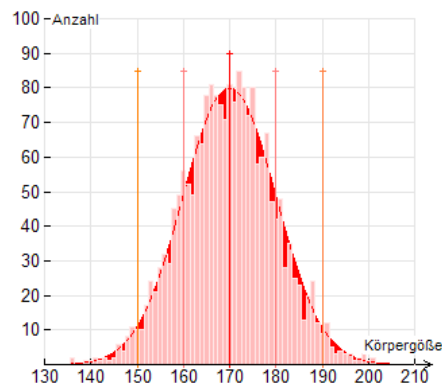
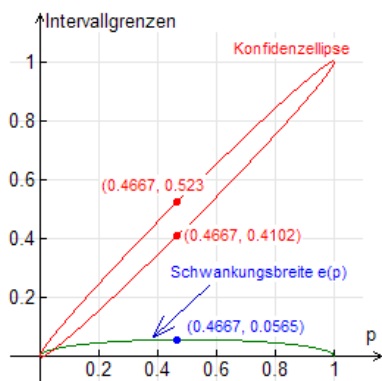
Elemente aus Statistik und Wahrscheinlichkeitsrechnung

Beispiele und Anregungen
zur Umsetzung im Unterricht mit TI-InterActive!
2008

Friedrich Tinhof

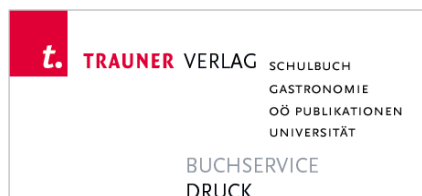
BHAK Eisenstadt

fritz.tinhof@t3oesterreich.at



„Some Mathematics becomes more important because technology requires it.
Some Mathematics becomes less important because technology replaces it.
Some Mathematics becomes possible because technology allows it.”

Bert Waits, Ohio State University (2000)



1 Empirische Verteilungen; Häufigkeitsverteilungen

Eine Grundgesamtheit ist durch eine Stichprobe hinsichtlich eines Merkmals zu untersuchen. Das Merkmal nimmt in dieser Stichprobe bestimmte Merkmalsausprägungen an. Für jede Merkmalsausprägung wird festgestellt, wie oft sie auftritt. Die Häufigkeit f_i (frequency) der i -ten Merkmalsausprägung wird ermittelt.

Beispiel: In einem Wohnblock wird die Anzahl der in einem Haushalt lebenden Personen, die Haushaltsgröße von 20 Haushalten erhoben. (Haushaltsgröße.tii)
 Es ergeben sich folgende Werte (Urliste): 2, 3, 5, 4, 6, 2, 2, 6, 2, 2, 2, 3, 1, 3, 3, 4, 3, 1, 7, 4

Merkmalsausprägung x_i Anzahl der im Haushalt lebenden Personen	Strichliste	absolute Häufigkeit f_i	relative Häufigkeit h_i	relative Häufigkeit in Prozent
1	//	2	0.1	10%
2	### /	6	0.3	30%
3	###	5	0.25	25%
4	///	3	0.15	15%
5	/	1	0.05	5%
6	//	2	0.1	10%
7	/	1	0.05	5%
	Summe	20	1	100%

$n = 20$ Stichprobenumfang

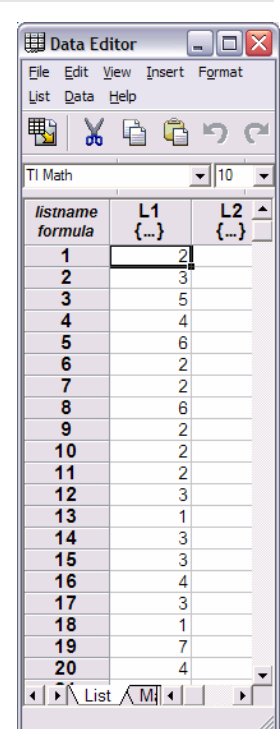
x_i Merkmalsausprägung, Haushaltsgröße, Anzahl der im Haushalt lebenden Personen




$x_i \in \{1, 2, 3, 4, 5, 6, 7\}$ $x_1 = 1; x_2 = 2; x_3 = 3; \dots; x_7 = 7$

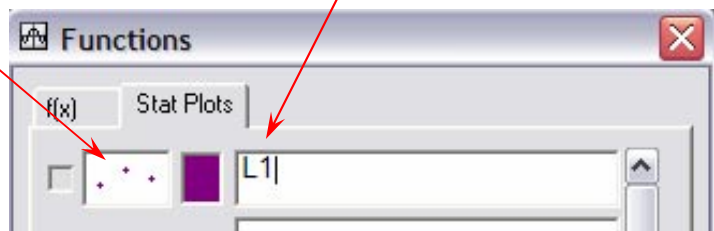
f_i absolute Häufigkeit (engl. frequency) der Haushaltsgröße x_i $\sum_{i=1}^7 f_i = n$

h_i relative Häufigkeit der Haushaltsgröße x_i ; $h_i = \frac{f_i}{20}$ $\sum_{i=1}^7 h_i = 1$

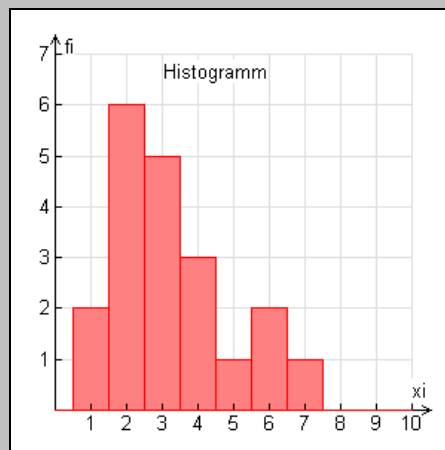
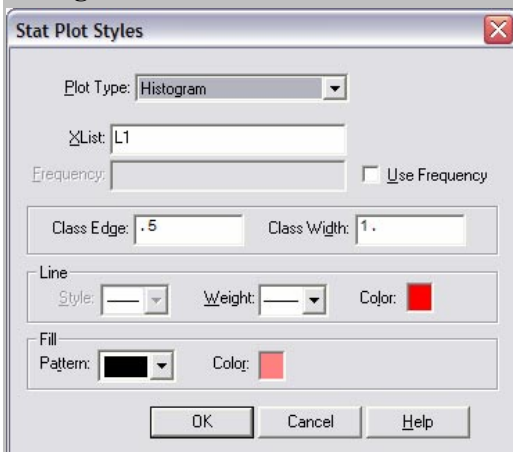
Darstellung der absoluten Häufigkeiten als Diagramm:



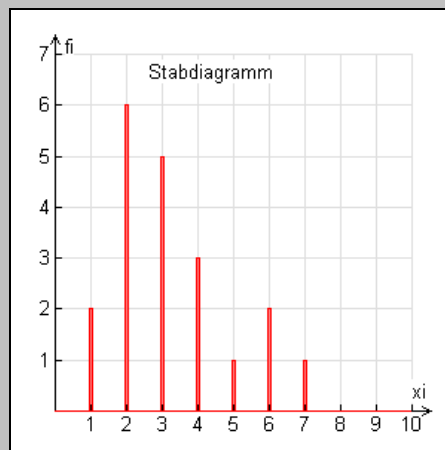
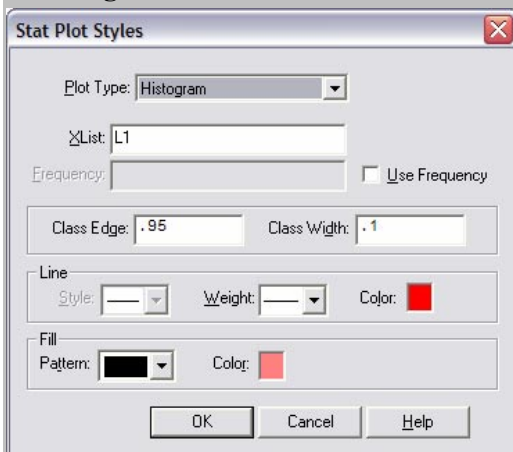
- Klicken Sie auf , um den **Data Editor** von TII zu öffnen.
 - Geben Sie die x_i -Werte in Liste L1 (und die f_i -Werte in Liste L2) ein.
 - Klicken Sie auf Ihr TII-Arbeitsblatt. Der Data Editor bleibt geöffnet.
 - Klicken Sie im **Data Editor** auf das Schaltzeichen , wählen Sie dann die statistische Grafik und geben Sie in das erste Eingabefenster den Namen der Liste L1 ein.
- Klicken Sie auf das Symbol , um die Grafik anzupassen und um sie zu formatieren.




Histogramm:



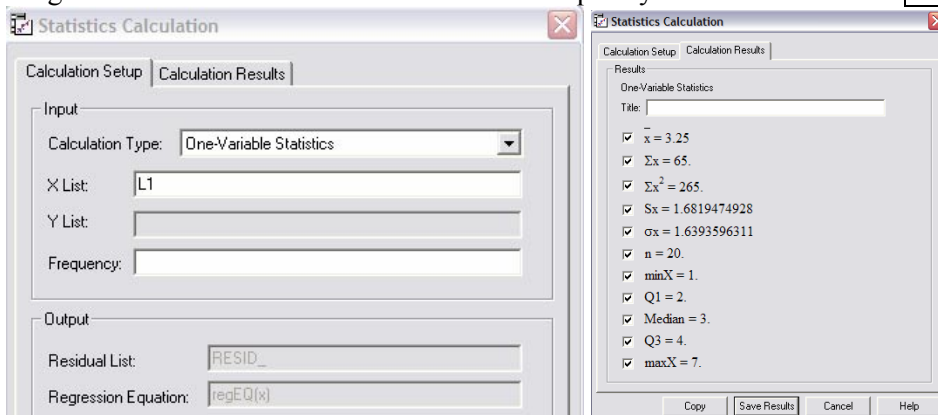
Stabdiagramm:



Statistische Kenngrößen:

Drücken Sie  und öffnen sie damit das **Stat Calculation Tool**.

Tragen Sie L1 für die Werte der X List ein. Frequency bleibt zunächst leer. **Calculate** & **Save Results**



`showstat()` =

xbar_	3.25
sumx_	65.
sumx2_	265.
sigmax_	1.63935963108
n_	20.
sx_	1.68194749277
minx_	1.
q1_	2.
med_	3.
q3_	4.
maxx_	7.

Mit der Eingabe **showstat()** in eine neue Mathbox erhalten Sie die Kenngrößen und deren Systembezeichnung.

Daten mit dem Zufallsgenerator (Pseudozufallszahlen):

Ganzzahlige Zufallszahlen erhält man mit **randint(untereGrenze, obereGrenze, Anzahl)**

Mit **randseed(code)** kann der Zufallsgenerator initialisiert werden, sodass alle Benutzer desselben Codes auch dieselben Zufallszahlen erhalten.

```
randint(1, 7, 20) = {2., 7., 5., 5., 7., 7., 3., 6., 1., 5., 6., 4., 1., 5., 2., 5., 3., 5., 7., 3.}
randint(1, 7, 20) = {3., 7., 7., 6., 2., 7., 3., 1., 7., 5., 5., 3., 3., 7., 5., 6., 7., 6., 3., 6.}
randseed(1234) = "Done"
randint(1, 7, 20) → L1 = {2., 3., 5., 4., 6., 2., 2., 6., 2., 2., 2., 3., 1., 3., 3., 4., 3., 1., 7., 4.}
```

Mit **STO→** L1 werden die ermittelten Zufallszahlen in Liste L1 gespeichert.

Sortieren der Listenelemente:

```
sortA(L1) = "Done"
L1 = {1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 6, 6, 7}
sortD(L1) = "Done"
L1 = {7., 6., 6., 5., 4., 4., 4., 3., 3., 3., 3., 3., 2., 2., 2., 2., 2., 2., 1., 1.}
```

Liniendiagramm:

Beim Liniendiagramm müssen die Koordinaten der Punkte in Listenform gegeben sein!

Da in L1 schon die Urliste steht, verwenden wir L2 und L3.

listname	L1	L2	L3	L4
formula	{...}	{...}	{...}	{...}
1	2	1	2	
2	3	2	6	
3	5	3	5	
4	4	4	3	
5	6	5	1	
6	2	6	2	
7	2	7	1	
8	6			

Stat Plot Styles

Plot Type: XY Line

XList: L2

YList: L3 Use Frequency

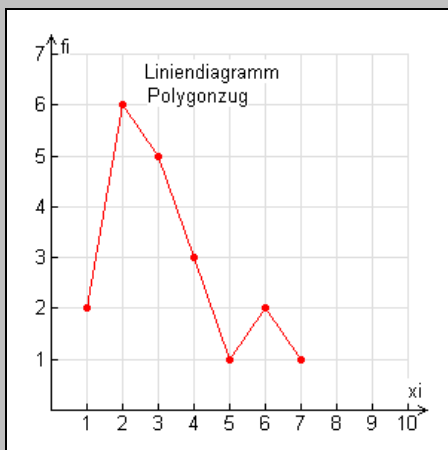
Line

Style: Weight: Color: ■

Mark

Symbol: Size:

OK Cancel Help



(haushaltsgröße.tii; randseed(1234))

Von besonderer Bedeutung für die folgenden Überlegungen ist auch die **Summenhäufigkeit** (kumulierte Häufigkeit) Fortsetzung des letzten Beispiels:

Im untersuchten Linzer Wohnblock wurde die Haushaltsgröße von 20 Haushalten erhoben. Die Auswertung der Daten erfolgte in Tabellenform.

- In wie vielen Haushalten leben bis zu maximal 2 Personen?
Wir müssen die Haushalte mit einer Person und die Haushalte mit 2 Personen addieren.

$$2 + 6 = f_1 + f_2 = \sum_{j=1}^2 f_j = 8 \quad \text{In 8 Haushalten leben maximal 2 Personen.}$$

- In wie vielen Haushalten leben bis zu maximal 3 Personen?
Wir müssen die Haushalte mit einer, zwei und die Haushalte mit drei Personen addieren.

$$2 + 6 + 5 = f_1 + f_2 + f_3 = \sum_{j=1}^3 f_j = 13 \quad \text{In 13 Haushalten leben maximal 3 Personen.}$$

- In wie vielen Haushalten leben bis zu maximal 6 Personen?
Wir müssen die Haushalte mit einer, zwei, drei, vier, fünf und die Haushalte mit sechs Personen addieren.

$$2 + 6 + 5 + 3 + 1 + 2 = f_1 + f_2 + f_3 + f_4 + f_5 + f_6 = \sum_{j=1}^6 f_j = 19$$

In 19 Haushalten leben maximal 6 Personen.

$$F_i = \sum_{j=1}^i f_j \quad \text{absolute Summenhäufigkeit (kumulierte Häufigkeit)}$$

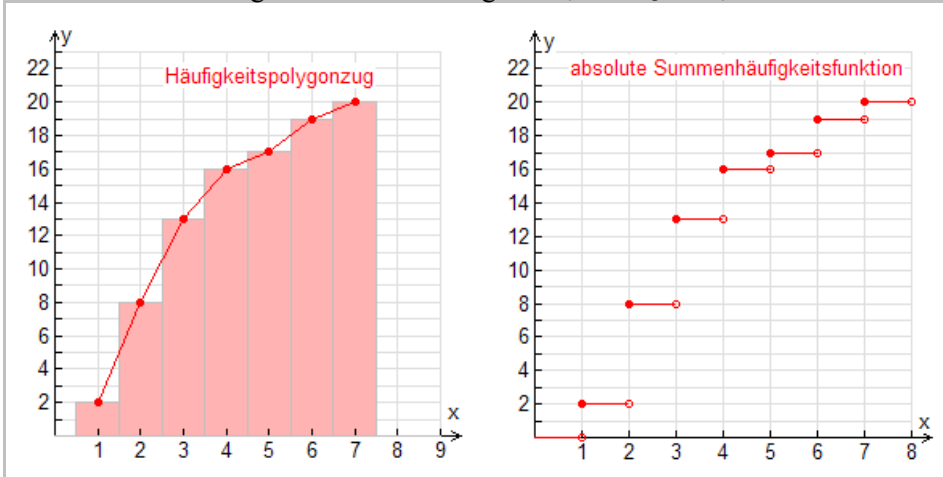
$$H_i = \sum_{j=1}^i h_j \quad \text{relative Summenhäufigkeit}$$

Fortsetzung des letzten Beispiels:

Darstellung in Tabellenform: i ist die Zeilennummer

i	x_i	f_i	h_i	absolute Summenhäufigkeit	relative Summenhäufigkeit
1	1	2	0.1	2	0.1
2	2	6	0.3	8	0.4
3	3	5	0.25	13	0.65
4	4	3	0.15	16	0.8
5	5	1	0.05	17	0.85
6	6	2	0.1	19	0.95
7	7	1	0.05	20	1
	Summe	20	1		

Grafische Darstellung der Summenhäufigkeit: (haushaltsgröße.tii)



Aus Summenhäufigkeiten können wieder die einzelnen Häufigkeiten ermittelt werden:

$$f_i = F_i - F_{i-1} \quad i = 1, 2, \dots k$$

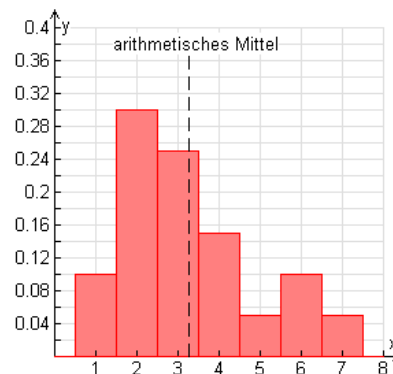
$$h_i = H_i - H_{i-1} \quad i = 1, 2, \dots k$$

Mit $F_0 = 0$ und $H_0 = 0$

Beispiel Fortsetzung: Die Tabelle gibt die Haushaltsgröße von 20 Haushalten in einem Wohnblock an. Wieviele Personen leben im Durchschnitt in einem Haushalt?

	L1	L2	L3	L4
i	x_i	f_i	h_i	$x_i \cdot h_i$
1	1	2	0.1	0.1
2	2	6	0.3	0.6
3	3	5	0.25	0.75
4	4	3	0.15	0.6
5	5	1	0.05	0.25
6	6	2	0.1	0.6
7	7	1	0.05	0.35
	Summe:	20	1	3.25

Durchschnittlich leben 3.25 Personen in einem Haushalt.



Median

Der Median darf ab ordinalskalierten Daten (Reihenfolge gegeben) berechnet werden.

Der Median teilt einen der Größe nach geordneten Datenbestand in zwei gleich große Hälften und wird von Ausreißern kaum beeinflusst.

Definition

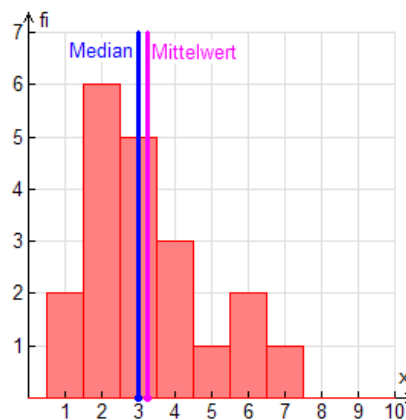
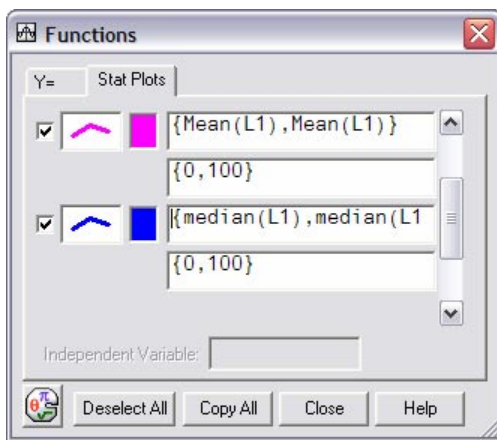
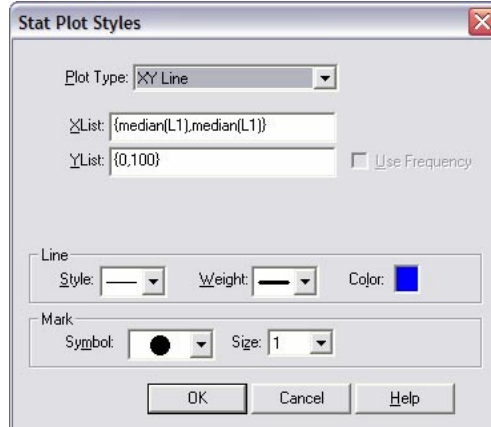
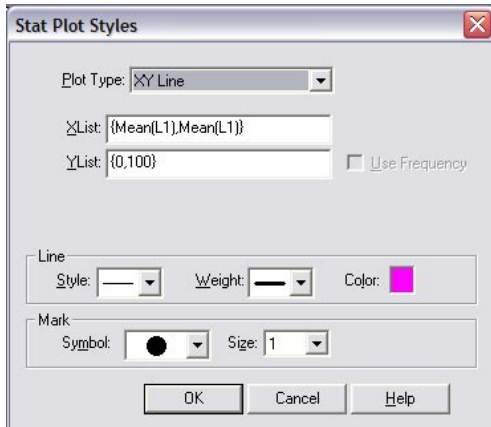
Der Median von Messwerten, die ihrer Größe nach geordnet sind, ist

- der Wert in der Mitte, bei ungerader Anzahl von Daten.
- bei gerader Anzahl von Daten das arithmetische Mittel der beiden Datenwerte in der Mitte.

Beispiel:

2, 3, 3, 4, 4, 5, 6, 7, 8 Median = 4
 2, 3, 3, 4, 4, 5, 6, 7, 8, 8 Median = (4 + 5)/2 = 4.5

Fortsetzung des Beispiels Haushaltsgröße
Median = 3



Quartile

Der Median teilt die der Größe nach geordneten Daten in eine obere und in eine untere Hälfte.

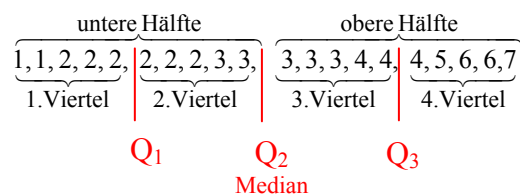
Das untere Quartil Q_1 ist der Median der unteren Datenhälfte.

Das obere Quartil Q_3 ist der Median der oberen Datenhälfte.

Das mittlere Quartil Q_2 ist der Median aller Daten.

Urliste: 2, 3, 5, 4, 6, 2, 2, 6, 2, 2, 2, 3, 1, 3, 3, 4, 3, 1, 7, 4

Der Größe nach geordnete Daten: 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 6, 6, 7



$$Q_1 = \frac{2+2}{2} = 2$$

$$Q_2 = \frac{3+3}{2} = 3$$

$$Q_3 = \frac{4+4}{2} = 4$$

randseed(1234) ⇒ "Done"
randint(1, 7, 20) → L 1

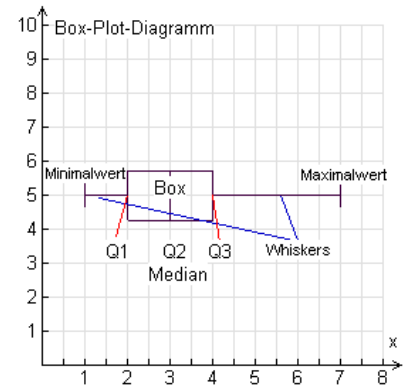
Die **Fünf-Punkte-Zusammenfassung** (nach *John Turkey*) einer Häufigkeitsverteilung besteht aus 5 Maßzahlen:

1. Minimalwert
2. unteres Quartil Q_1
3. Median Q_2
4. oberes Quartil Q_3
5. Maximalwert

Ein **Box-Plot-Diagramm** ist eine grafische Veranschaulichung der Fünf-Punkte-Zusammenfassung.

Die Box erstreckt sich über das Intervall $[Q_1 ; Q_3]$, den **Quartilsabstand**. Der Median wird in der Box durch einen Strich markiert.

Zwei Linien (whiskers, Schnurrhaare einer Katze) außerhalb der Box gehen bis zum Minimalwert bzw. Maximalwert.



Fortsetzung des Beispiels Haushaltsgröße. (haushaltsgröße.tii/pdf)

Die Datenwerte liegen in Liste L1 vor.

Lösung mit TII

Klicken Sie auf den Pfeil neben dem Schaltsymbol für Grafik.

Wählen Sie die statistische Darstellung.


Im Eingabefenster für statistische Daten geben Sie für

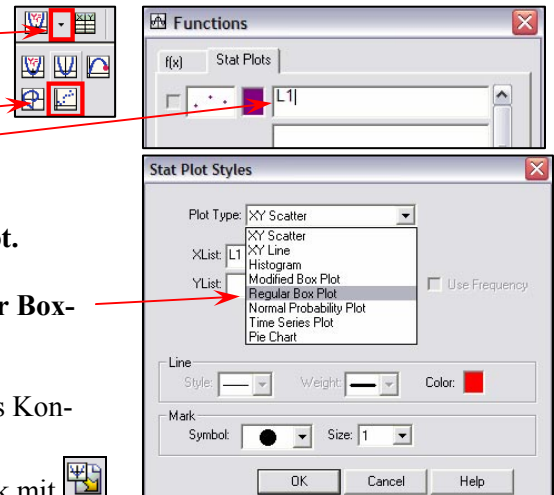
XList L1 ein.

Ändern Sie **Weight** auf die **mittlere Dicke** und die Farbe auf **rot**.

Aus der Auswahlliste für den Plot Type wählen Sie den **Regular Box-Plot**.

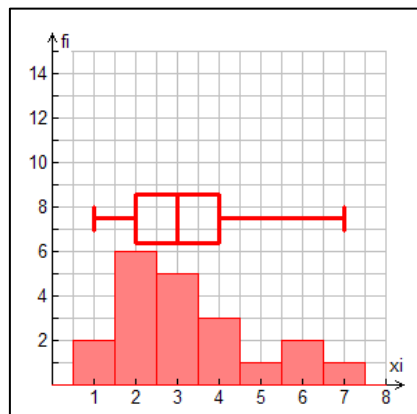
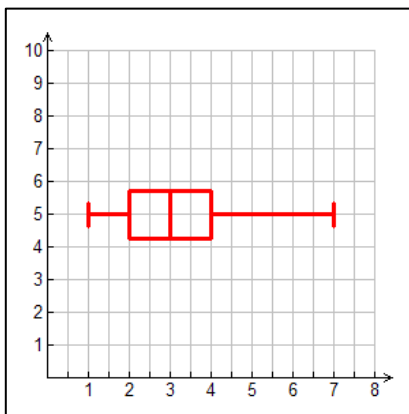
Bestätigen Sie Ihre Einstellungen mit **OK** und aktivieren Sie das Kontrollkästchen vor der eingegebenen Liste.

Passen Sie die Fenstereinstellungen an und fügen Sie Ihre Grafik mit  in das aktuelle Arbeitsblatt ein

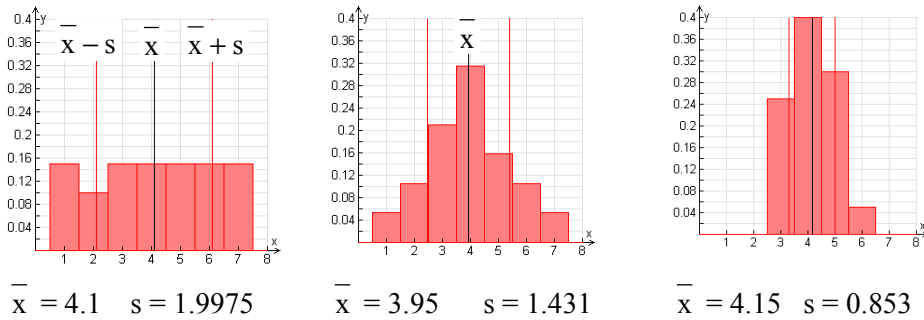


Anmerkung: Das **Modified BoxPlot** stellt Ausreißer als Einzelpunkte dar.

Verschiedene Darstellungsformen können auch kombiniert werden.



Veranschaulichung und Interpretation der Standardabweichung



Die Standardabweichung s ist umso größer, je mehr die Daten gestreut sind.

Bei umfangreichen, annähernd „normalverteilten“ Daten gilt:
 ca. 70% der Werte liegen im einfachen Streuintervall $[\bar{x} - s ; \bar{x} + s]$
 ca. 95% der Werte liegen in zweifachen Streuintervall $[\bar{x} - 2 \cdot s ; \bar{x} + 2 \cdot s]$

Beispiel:
 Die Körpergröße einer Population ($n = 2000$) sei angenähert normalverteilt. (koerpergr1.tii/pdf; 2219)
 Der Mittelwert sei $\bar{x} = 170$ cm, die Standardabweichung sei $s = 10$ cm.

Es gilt näherungsweise:

68% (1360 Personen) der Population sind zwischen 160 cm und 180 cm groß.

95% (1900 Personen) der Population sind zwischen 150 cm und 190 cm groß.

Eingaben:

Initialisierung:

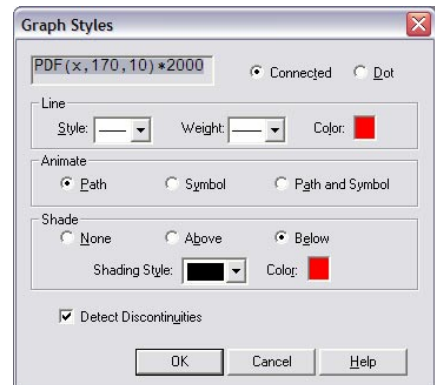
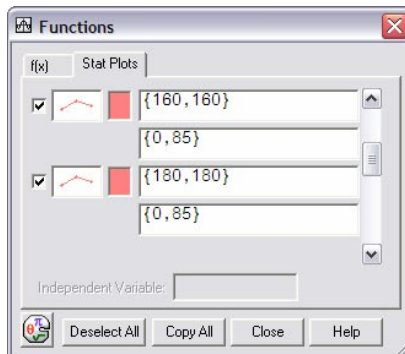
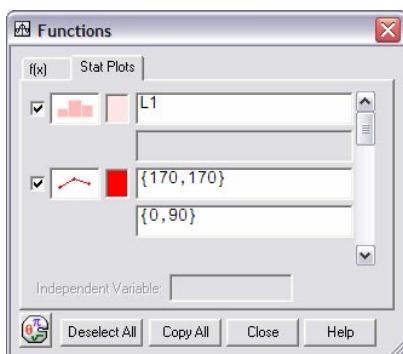
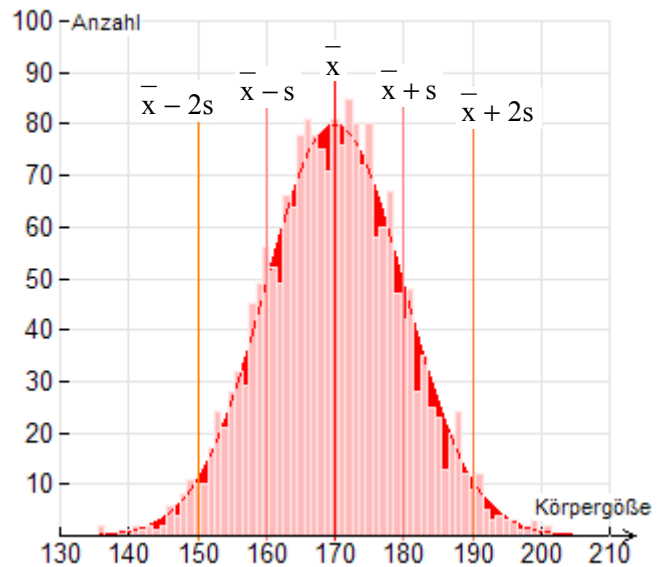
`randseed(2219) => "Done"`

2000 normalverteilte Zufallszahlen in L1

($\mu=170$; $\sigma = 10$): `randnorm(μ,σ,n)`

`L1 := randnorm(170, 10, 2000)`

`f1(x): normalPDF(x,170,10)*2000`



2 Zweidimensionale Verteilungen; Regressionsrechnung

Wir haben bisher nur eindimensionale Häufigkeitsverteilungen untersucht. In der Praxis findet man aber oft einen Zusammenhang zwischen zwei Variablen X und Y.

Regressionsanalyse, Trendlinie

Mit der Regressionsanalyse wird versucht, den Zusammenhang von **quantitativen Merkmalen** in Form einer mathematischen Funktion anzugeben.

Beispiel:

Von 5 zufällig gewählten Personen wurden Körpergröße und Körpermasse (Gewicht) gemessen.

Gibt es einen Zusammenhang (Trend) zwischen Körpergröße und Körpermasse? ([Körpergröße5pers.tii/pdf](#))

Name	Körpergröße in cm	Körpermasse in kg
Antonia	165	56.3
Berta	178	71.5
Claudia	171.5	70.3
Dorothea	160.3	63.4
Erika	175	67.3

Die Datenpaare $(x_i | y_i)$ werden als Punkte im **Streudiagramm** (engl. Scatter Plot) in einem kartesischen Koordinatensystem dargestellt. Die dargestellten Punkte bilden eine **Punktewolke**.

Gesucht ist nun eine **Trendlinie**, dh. zunächst eine Kurve, die den Zusammenhang zwischen Körpergröße und Körpermasse am besten wiedergibt.

Die Frage ist nur, welche Kurve ist „die Beste“?

Der einfachste Typ einer Näherungskurve ist die Gerade mit der Gleichung $y = a \cdot x + b$.

Die Einflussgröße x (unabhängige Variable, Körpergröße) bedingt die Werte der Zielgröße y (abhängige Variable, Körpermasse).

Zur Anpassung der Näherungskurve an das Streudiagramm wird üblicherweise die

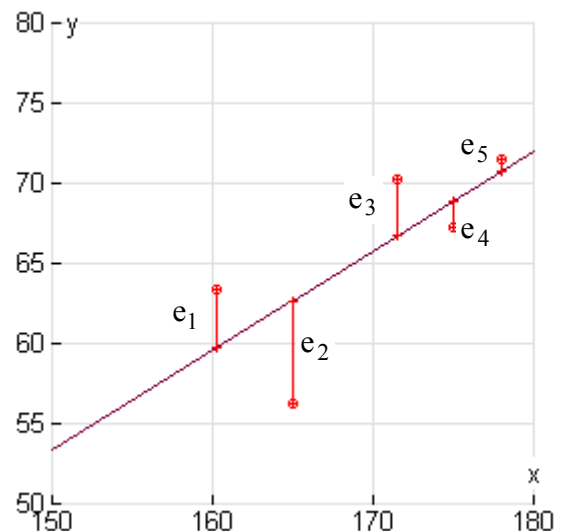
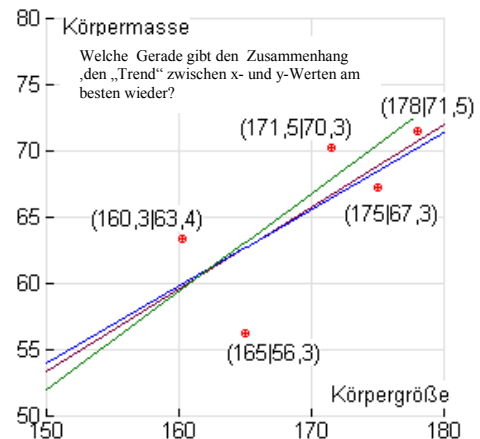
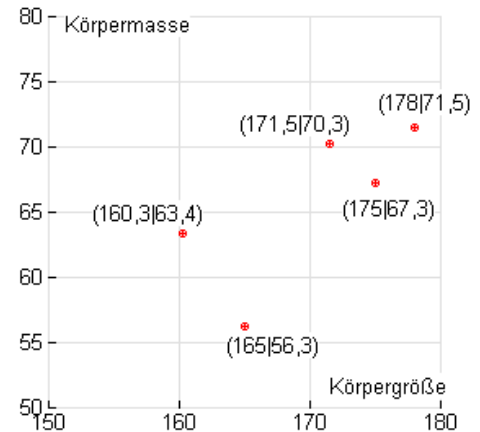
Methode der kleinsten Quadrate verwendet.

Wir passen eine Gerade mit der Gleichung $y = a \cdot x + b$ so an das Streudiagramm an, dass die Summe der Quadrate der vertikalen Fehler (Residuen e_i) minimal wird.

Der Fehler zwischen dem berechneten y-Wert auf der Regressionslinie und dem gemessenen y-Wert für das Wertepaar $(x_i | y_i)$ ist das **Residuum** e_i .

$$e_i = \underbrace{y_i}_{\text{Messwert}} - \underbrace{(a \cdot x_i + b)}_{\text{Modellwert}} \quad i = 1, 2, 3, \dots, n$$

Die Residuen sind die Abstände zwischen den gemessenen Werten y_i und den Modellwerten (Punkten auf der Regressionslinie).



Die **Quadratsumme F aller Fehler** (Residuen) der n gemessenen Wertepaare $(x_i|y_i)$ beträgt:

$$F(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left(\underbrace{(a \cdot x_i + b)}_{y(x_i)} - y_i \right)^2 = \sum_{i=1}^n \left(\underbrace{\frac{y(x_i)}{\text{Modellwert}} - \frac{y_i}{\text{Messwert}}}_{\text{Summe der Fehlerquadrate}} \right)^2 \Rightarrow \text{Minimum}$$

Die Modellparameter a und b sind so zu berechnen, dass **F(a,b) minimal** wird.

Händisch ist dies sehr rechenintensiv.

Mit CAS ist die Lösung der Extremwertaufgabe einfach zu berechnen.

Lösung mit TII

L1	L2
165	56.3
178	71.5
171.5	70.3
160.3	63.4
175	67.3

$$y(x) := a \cdot x + b = \text{"Done"}$$

Ansatz der Regressionsgeraden

$$n := \dim(L1) = 5$$

$$F(a, b) := \sum_{i=1}^n \left((y(L1_{[i]}) - L2_{[i]})^2 \right) = \text{"Done"}$$

$$F(a, b) = 144642 \cdot a^2 + a \cdot (1699.6b - 112027.) + 5 \cdot b^2 - 657.6b + 21772.9$$

partielle Ableitung nach a: $F_a = 0$ -> gl1 (Gleichung1)

$$\frac{d}{da} (F(a, b)) = 289284.68a + 1699.6b - 112026.94$$

$$\text{ans} = 0 \rightarrow \text{gl1} = 289285 \cdot a + 1699.6b - 112027. = 0$$

partielle Ableitung nach b: $F_b = 0$ -> gl2 (Gleichung2)

$$\frac{d}{db} (F(a, b)) = 10 \cdot b + 1699.6a - 657.6$$

$$\text{ans} = 0 \rightarrow \text{gl2} = 1699.6a + 10 \cdot b - 657.6 = 0$$

Lösen des Gleichungssystems gl1,gl2

$$\text{solve}(\text{gl1}, a) \Rightarrow a = -.005875 \cdot (b - 65.9137)$$

$$\text{right}(\text{ans}) \rightarrow a = -.005875 \cdot (b - 65.9137)$$

$$\text{solve}(\text{gl2}, b) \Rightarrow b = -39.7899$$

$$\text{right}(\text{ans}) \rightarrow b = -39.7899$$

$$a = .621028$$

$$y(x) = .621028 \cdot x - 39.7899 \quad \text{Regressionsgleichung}$$

Auf analoge Weise kann auch Regressionsrechnung mit Funktionen höheren Grades betrieben werden.
(methode_d_kl_Quadr.tii/pdf)

Anmerkung: Allgemein gültige Berechnungsformeln zur Regression finden Sie in einschlägigen Büchern.
Auf die Möglichkeit Regressionslinien mit Hilfe von Matrizen zu berechnen wird hier nicht eingegangen.

Wünschenswert ist jetzt noch eine Aussage über den Zusammenhang der statistischen Daten in Form einer Kennzahl genauer auszudrücken.

Dafür definieren wir den Pearson'schen Korrelationskoeffizienten

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

Mittelwert der x-Koordinaten der gegebenen Daten

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^n y_i$$

Mittelwert der y-Koordinaten der gegebenen Daten

$(\bar{x} | \bar{y})$ entspricht dem Schwerpunkt der Punktwolke

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Kovarianz

positiv, wenn die Daten von links unten nach rechts oben liegen

negativ, wenn die Daten von links oben nach rechts unten liegen

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Korrelationskoeffizient (Pearson)

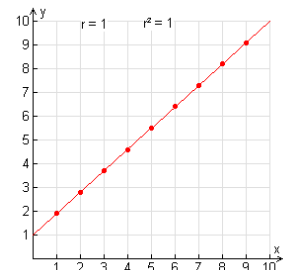
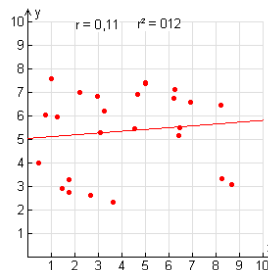
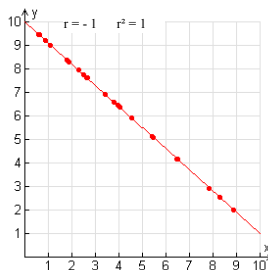
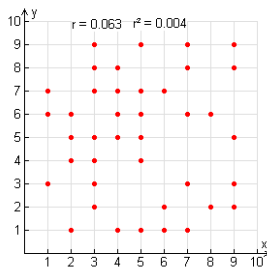
Die Kovarianz wird durch die Standardabweichungen s_x und s_y dividiert.

Durch diese Normierung gilt für r : $-1 \leq r \leq 1$.

$$R^2 = r^2 = B$$

Bestimmtheitsmaß $0 \leq r^2 \leq 1$

Der Pearson'sche Korrelationskoeffizient r ist auch ein Maß für die Linearität des Zusammenhanges zweier Merkmale. Er ist daher für nichtlineare Zusammenhänge nicht geeignet.



Wegen des großen Rechenaufwandes wird die Trendlinie in Folge ausschließlich mit Technologieunterstützung berechnet.




T³ ÖSTERREICH

Trendlinien mit Hilfsmitteln berechnet

Der GTR (TI-83/TI-84) bietet 11 Regressionsvarianten, die es ermöglichen schnell und einfach die Gleichung einer Regressionskurve zu berechnen. TII und andere Softwareprodukte bieten noch mehr Möglichkeiten für die Regression.

Berechnung der Trendlinie mit TI InterActive! (Fortsetzung des Beispiels)

- ☞ Geben Sie im ersten Schritt die gegebenen Daten in Listen **L1** und **L2** ein.
- ☞ Markieren Sie die Listen mit den Daten.
- ☞ Öffnen Sie das **Stat Calculation Tool** mit .
- ☞ Eingabe in das Fenster der **Statistics Calculation**:
Calculation Type: aus der Liste **Linear Regression (ax+b)** wählen

listname	L1	L2
1	165	56.3
2	178	71.5
3	171.5	70.3
4	160.3	63.4
5	175	67.3

XList: L1

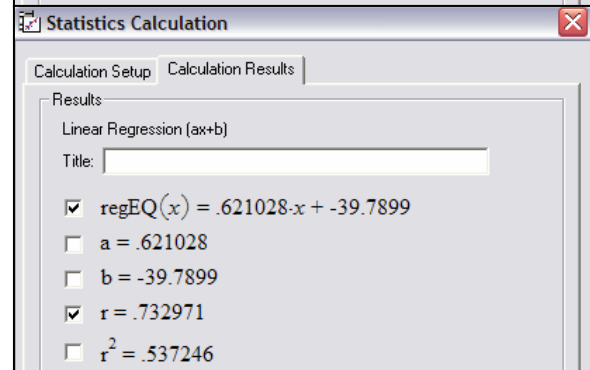
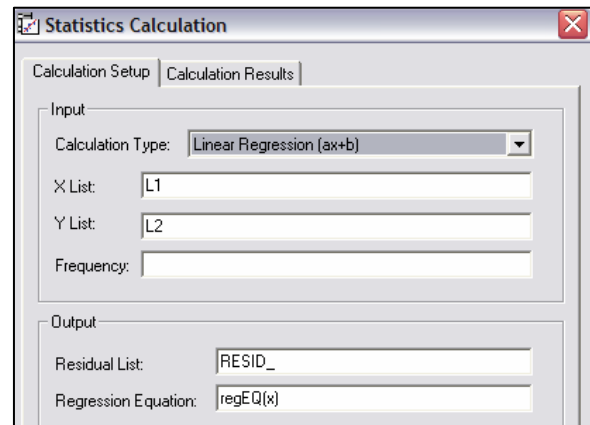
YList: L2

Residual List: Listenname frei wählbar


Regression Equation: Name der Funktion(x)

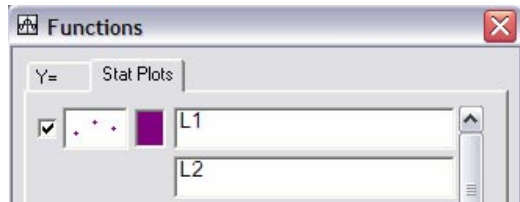
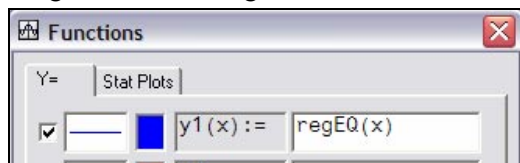
Calculate

- ☞ Markieren () Sie Rechenergebnisse, die auf Ihrem Arbeitsblatt dargestellt werden sollen.
- ☞ Klicken Sie auf **Save Results**, um die Ergebnisse der Regressionsrechnung ins Arbeitsblatt zu übernehmen.



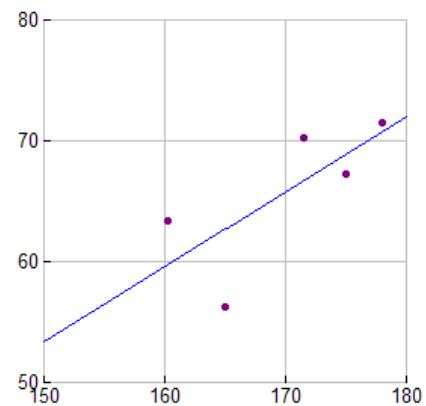
Grafik

- ☞ Klicken Sie auf .
- ☞ Eingaben in das Eingabefenster:



- ☞ Fenstereinstellungen:

Xmin:	150.
Xmax:	180.
Xscale:	10.
Ymin:	50.
Ymax:	80.
Yscale:	10.



Einfügen der **Grafik** mit  und des **Data Editors** mit  in Ihr Arbeitsblatt.

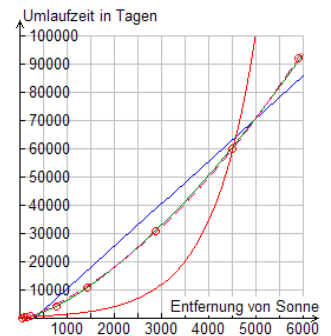
Das Arbeitsblatt muss unter Umständen noch im Layout verbessert werden. [\(regressionsrechnung.tii\)](#)

Aufgaben

Ü1) Das 3. Gesetz von Kepler

Die Planeten bewegen sich auf elliptischen Bahnen um die Sonne. Es ist offensichtlich, dass der Abstand von der Sonne einen Einfluss auf die Umlaufzeit der Planeten hat. Die Frage ist nur, ob dieser Zusammenhang linear, quadratisch, exponentiell oder von einer ganz anderen Form ist. Untersuchen Sie mit Hilfe der Regressionsrechnung den funktionalen Zusammenhang zwischen Bahnradius (genauer: großer Ellipsenhalbachse) und Umlaufzeit. Vergleichen Sie die Regressionstypen linear, quadratisch, exponentiell und potenziell. Berechnen Sie r^2 .
Anmerkung: Vergrößern Sie in Ihrer Grafik den Bereich für Merkur, Venus, Erde und Mars. Welche Regression modelliert in dieser Vergrößerung die Punktwolke besser?

	Mittlerer Bahnradius in 10^6 km	Umlaufzeit
Sonne	—	—
Merkur	57,87	88,00 d
Venus	108,14	225,00 d
Erde	149,60	365,25 d
Mars	227,80	687,00 d
Jupiter	777,84	11,9 J
Saturn	1 426,10	29,5 J
Uranus	2 867,83	84,0 J
Neptun	4 493,65	164,8 J
Pluto	5 899,04	251,9 J

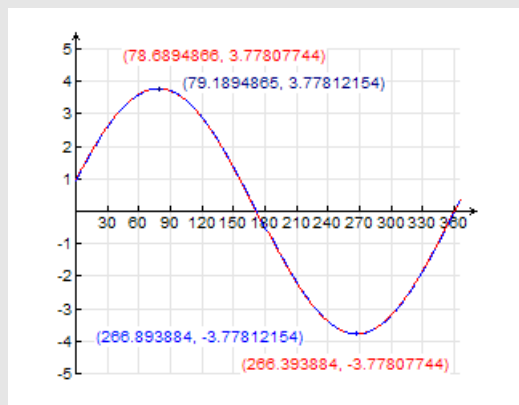


[\(Kepler3.tii/pdf\)](#)

Ü2) Sinus Regression

Die Tabelle enthält die Tageslänge in Minuten im nördlichen Burgenland, gemessen an bestimmten Kalendertagen. Berechnen Sie eine Funktion, die die Punkte optimal annähert. An welchem Tag ist die Tageslänge maximal? An welchem Tag ist die Zunahme der Tageslänge maximal?

Tag	Tageslänge
1	506
32	569
60	660
91	769
121	869
151	944
213	902
274	703
335	521
365	505



[\(SinusTaglaeng.tii/pdf; vergleiche: 366Tage.tiipdf\)](#)



T³ ÖSTERREICH

2 Wahrscheinlichkeitsrechnung

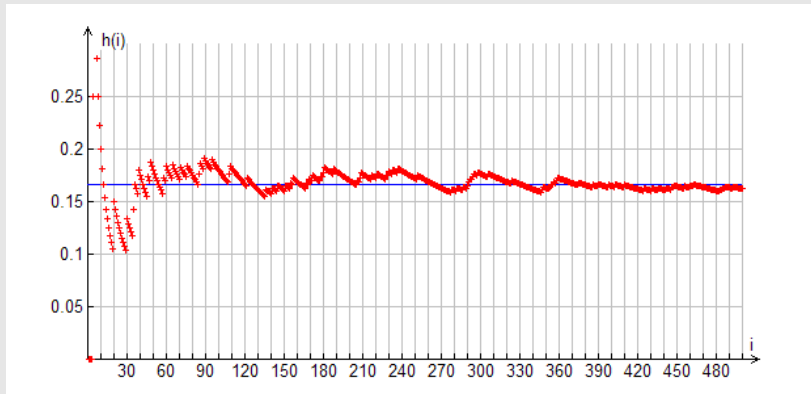
Historischer Wahrscheinlichkeitsbegriff: $P(A) \approx h(n)$ für große n

Die **relative Häufigkeit $h(n)$** stellt für eine hinreichend große Anzahl von Durchführungen des Zufallsexperiments einen **Schätzwert** für die gesuchte **Wahrscheinlichkeit P** dar.

Beispiel: Darstellung der relativen Häufigkeit $h(i)$ mit TI InterActive!. (WerfenWuerfel.tii)

Die Darstellung wurde durch Verwendung eines kleinen Programms automatisiert.

Untersuchtes Ereignis: Augenzahl 6 bei 500 Würfeln



```

Define Wuerfel(a,n) = Func
local a,i
::seq(a,a,1,n)→L1
::randint(1,6,n)→L2
::for i,1,n
::if L2[i]=x Then
::1→L3[i]
::Else
::0→L3[i]
::EndIf
::EndFor
::cumsum(L3)→L4
:L4/L1→L5
EndFunc
= "Done"
    
```

Die Häufigkeit $h(i)$ „stabilisiert“ sich in der Nähe der Wahrscheinlichkeit $P \approx 1/6$.

Es stellt sich die Frage: Kann man rel. Häufigkeiten $h(i)$ vorhersagen oder wenigstens einschränken?

Beispiel: In einer Klasse der BHAK Eisenstadt befinden sich 25 Schüler.

- Wie groß ist die Wahrscheinlichkeit, dass an einem Tag mehr als ein Schüler Geburtstag hat?
- Wie groß ist die Wahrscheinlichkeit, dass allgemein bei n Personen an einem Tag mehr als eine Person Geburtstag hat?
- Ab wie vielen Personen ist es wahrscheinlicher, dass mehrere Personen am selben Tag Geburtstag haben, als dass alle Personen an verschiedenen Tagen Geburtstag haben?

Bei der Lösung der Aufgabe nehmen wir an, dass alle 365 Tage eines Jahres gleichwertig sind und wir vernachlässigen Schaltjahre. (Geburtsstagsproblem.tii)

\bar{A} ... mehrere Personen haben an einem Tag Geburtstag

\bar{A} ... alle Personen haben an verschiedenen Tagen Geburtstag

Zahl der Personen n	Wahrscheinlichkeit, dass n Personen an verschiedenen Tagen Geburtstag feiern. $P(\bar{A})$	Wahrscheinlichkeit, dass bei n Personen mehrere Personen an einem Tag Geburtstag haben. $P(A) = 1 - P(\bar{A})$
1	$\frac{365}{365} = \frac{366-1}{365} = 1$	$1 - \frac{365}{365} = 0$
2	$\frac{365}{365} \cdot \frac{364}{365} = \frac{365}{365} \cdot \frac{366-2}{365} \approx 0,997$	$1 - \frac{365}{365} \cdot \frac{364}{365} \approx 0,003$
3	$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{366-3+1}{365} = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{366-3}{365}$	$1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \approx 0,008$
4	$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{366-4}{365} = 0,984$	$1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \approx 0,016$
...

Binomialverteilung

Die Binomialverteilung (Bernoulli-Verteilung) ist wahrscheinlich die wichtigste diskrete Wahrscheinlichkeitsverteilung.

Grundlage ist ein Zufallsexperiment mit zwei möglichen, einander ausschließenden Ergebnissen A und \bar{A} .

$P(A) = p$ **Erfolgswahrscheinlichkeit**

$P(\bar{A}) = 1 - p = q$ **Misserfolgswahrscheinlichkeit**

Das Zufallsexperiment wird n -mal durchgeführt, wobei die Durchführungen unabhängig voneinander sein sollen (Ziehen mit Zurücklegen).

Untersucht wird die **Anzahl der Erfolge x bei n unabhängigen Durchführungen** des Zufallsexperiments.

Beispiel:

In einer Urne befinden sich 15 Kugeln. Davon sind **7 rot** und **8 Kugeln blau**. Es werden 3 Kugeln mit Zurücklegen entnommen. Wie groß ist die Wahrscheinlichkeit keine, eine, zwei oder drei rote Kugeln zu ziehen?

$N = 15$ **Umfang der Grundmenge**

$M = 7$ **Anzahl der Elemente mit bestimmter Eigenschaft (hier: Farbe der Kugel ist rot).**

$p = \frac{M}{N} = \frac{7}{15}$ **Erfolgswahrscheinlichkeit p**

$q = 1 - \frac{M}{N} = \frac{8}{15}$ **Misserfolgswahrscheinlichkeit $q = 1 - p$**

$n = 3$ **Umfang der Stichprobe (mit Zurücklegen)**

$X =$ Anzahl der roten Kugeln in der Stichprobe, $x_i \in \{0, 1, 2, 3\}$

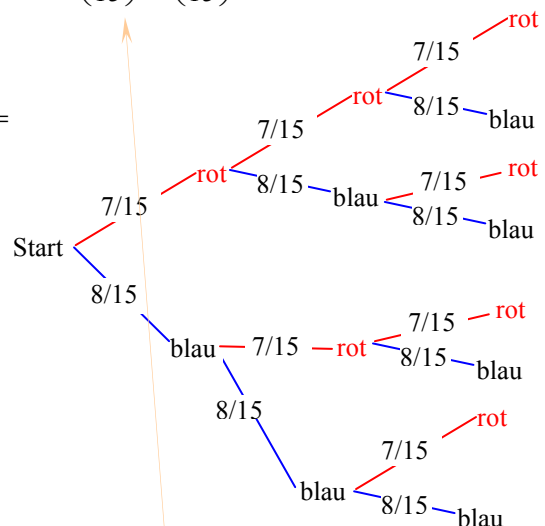
$$P(X = 0) = P(\text{nie rot}) = P(b_1 \cap b_2 \cap b_3) = \frac{8}{15} \cdot \frac{8}{15} \cdot \frac{8}{15} = 0.517 \left(= 1 \cdot \left(\frac{7}{15}\right)^0 \cdot \left(\frac{8}{15}\right)^3 \right)$$

$$\begin{aligned} P(X = 1) &= P(\text{einmal rot und zweimal nicht rot}) = \\ &= P((r_1 \cap b_2 \cap b_3) \cup (b_1 \cap r_2 \cap b_3) \cup (b_1 \cap b_2 \cap r_3)) = \\ &= P(r_1 \cap b_2 \cap b_3) + P(b_1 \cap r_2 \cap b_3) + P(b_1 \cap b_2 \cap r_3) = \\ &= \frac{7}{15} \cdot \frac{8}{15} \cdot \frac{8}{15} + \frac{8}{15} \cdot \frac{7}{15} \cdot \frac{8}{15} + \frac{8}{15} \cdot \frac{8}{15} \cdot \frac{7}{15} = \\ &= 3 \cdot \left(\frac{7}{15}\right)^1 \cdot \left(\frac{8}{15}\right)^2 = 0.3982 \end{aligned}$$

$$\begin{aligned} P(X = 2) &= P(\text{zweimal rot und einmal nicht rot}) = \\ &= P((r_1 \cap r_2 \cap b_3) \cup (r_1 \cap r_2 \cap b_3) \cup (r_1 \cap r_2 \cap b_3)) = \\ &= P(r_1 \cap r_2 \cap b_3) + P(r_1 \cap b_2 \cap r_3) + P(b_1 \cap r_2 \cap r_3) = \\ &= \frac{7}{15} \cdot \frac{7}{15} \cdot \frac{8}{15} + \frac{7}{15} \cdot \frac{8}{15} \cdot \frac{7}{15} + \frac{8}{15} \cdot \frac{7}{15} \cdot \frac{7}{15} = \\ &= 3 \cdot \left(\frac{7}{15}\right)^2 \cdot \left(\frac{8}{15}\right)^1 = 0.3484 \end{aligned}$$

$$P(X = 3) = P(\text{dreimal rot}) = P(r_1 \cap r_2 \cap r_3) = \frac{7}{15} \cdot \frac{7}{15} \cdot \frac{7}{15} = 0.1016 = 1 \cdot \left(\frac{7}{15}\right)^3 \cdot \left(\frac{8}{15}\right)^0$$

Summe: $P = 1$



Die Lösung der Aufgabe anhand eines Ereignisbaumes ist eher mühsam. Für große n (z.B. 100 Kugeln werden gezogen) ist diese Vorgangsweise sogar fast undurchführbar. Eine Verallgemeinerung der Vorgangsweise ist angebracht.

Zusammenfassung: p = Erfolgswahrscheinlichkeit ; $q = 1 - p$ = Misserfolgswahrscheinlichkeit

$$P(X = \text{Anzahl der Erfolge}) = \text{Anzahl der Pfade} \cdot p^{\text{Anzahl der Erfolge}} \cdot q^{\text{Anzahl der Misserfolge}}$$

Die Anzahl der Pfade lassen sich mit Kombinatorik berechnen:

Wieviele Möglichkeiten gibt z.B. es aus 3 freien Plätzen 2 Plätze für p auszuwählen?

$$\binom{3}{2} = 3 \text{ Möglichkeiten} \quad (\text{ppq oder pqp oder qpp}) \quad (\text{siehe Seite 53})$$

Wieviele Möglichkeiten gibt es aus n freien Plätzen x Plätze für p auszuwählen?

$$\binom{n}{x} \text{ Möglichkeiten} \quad \binom{7}{2}$$

Diese Zahl gibt an, wie viele Pfade es gibt, bei n **Versuchen**, x **Erfolge** zu erhalten.

Binomialverteilung $B(n; p)$

$x \in \{0, 1, 2, 3, \dots, n\}$; diskret

$$f(x) = P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \quad \sum_{i=1}^n P(X = x_i) = 1 \quad \text{Wahrscheinlichkeitsfunkt.}$$

$$F(x_k) = P(X \leq x_k) = P(X = 0) + P(X = 1) + \dots + P(X = x_k) = \sum_{i=1}^k P(X = x_i) \quad \text{Verteilungsfunktion}$$

$$\mu = E(X) = n \cdot p \quad \text{Erwartungswert}$$

$$\sigma = \sqrt{V(X)} = \sqrt{n \cdot p \cdot (1-p)} \quad \text{Standardabweichung}$$

Verwendung von Technologie

Fortsetzung des letzten Beispiels: In einer Urne befinden sich 15 Kugeln. Davon sind 7 rot und 8 Kugeln blau. Es werden 3 Kugeln mit Zurücklegen entnommen. Wie groß ist die Wahrscheinlichkeit keine, eine, zwei oder drei rote Kugeln zu ziehen?

Bei TI-Produkten gibt es einheitliche Funktionen zur Berechnung der Binomialverteilung:

$$f = P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \quad \text{entspricht die Funktion } \mathbf{binompdf(n,p,x)}$$

$$F(x) = P(X \leq x) \quad \text{entspricht die Funktion } \mathbf{binomcdf(n,p,x)}.$$

Dabei steht die Abkürzung **pdf** für **probability distribution/density function** (Dichtefunktion) und **cdf** für **cumulative distribution/density function** (Summenfunktion, Verteilungsfunktion).

Die Berechnung ist mit TI bis $n = 999999$ möglich!


TI InterActive! (B3_7_15.tii)

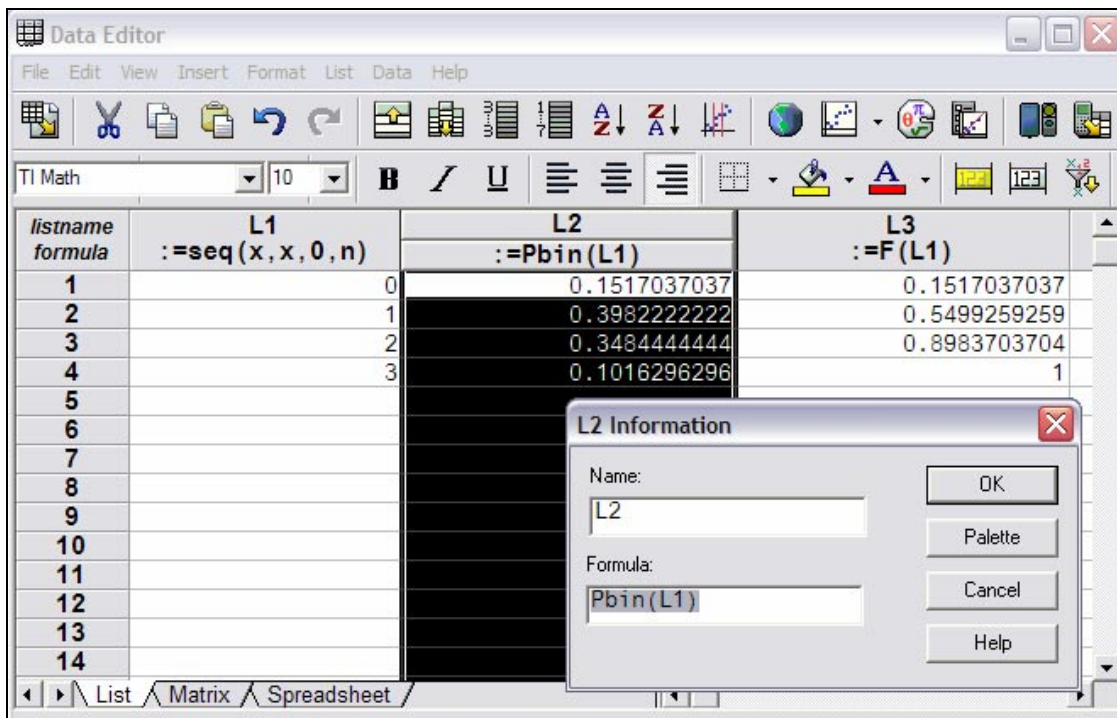
Definieren Sie im neuen Arbeitsblatt zunächst die vorliegenden Parameter n und p .


Es ist auch zweckmäßig (n und p ändern sich nicht) die verwendeten Funktionen vereinfacht neu zu definieren. Danach können Sie die gefragten Wahrscheinlichkeiten durch Einsetzen der $x_i \in \{0, 1, 2, \dots, n\}$ berechnen und die Ausgabe nach Wunsch formatieren.

Für $P(U \leq X \leq O)$ verwenden Sie $\mathbf{P(U \leq X \leq O) = F(O) - F(U-1)}$; mit U und $O \in \{0, 1, 2, \dots, n\}$

$n := 3$	Stichprobenumfang	
$p := \frac{7}{15}$	Erfolgswahrscheinlichkeit	$m := n \cdot p \quad s := \sqrt{n \cdot p \cdot (1 - p)}$
$P_{bin}(x) := \text{binomPDF}(n, p, x) \Rightarrow \text{"Done"}$		
$F(x) := \text{binomCDF}(n, p, x) \Rightarrow \text{"Done"}$		
$\sigma = .864099 \quad \mu = 1.4$		
$P(X = 0) = P_{bin}(0) = .151704 = \text{binompdf}(n, p, 0) = .151704$		
$P(X = 1) = P_{bin}(1) = .398222$		
$P(X = 2) = P_{bin}(2) = .348444$		
$P(X = 3) = P_{bin}(3) = .10163$		
$[P(X = 4) = P_{bin}(4) = 0. \quad \text{unmögliches Ereignis!}]$		
$P(X < 2) = P(X \leq 1) = P(X = 0) + P(X = 1) = F(1) = .549926$		
$P(1 < X \leq 3) = P(X = 2) + P(X = 3) = 1 - P(X \leq 1) = F(3) - F(1) = .450074$		

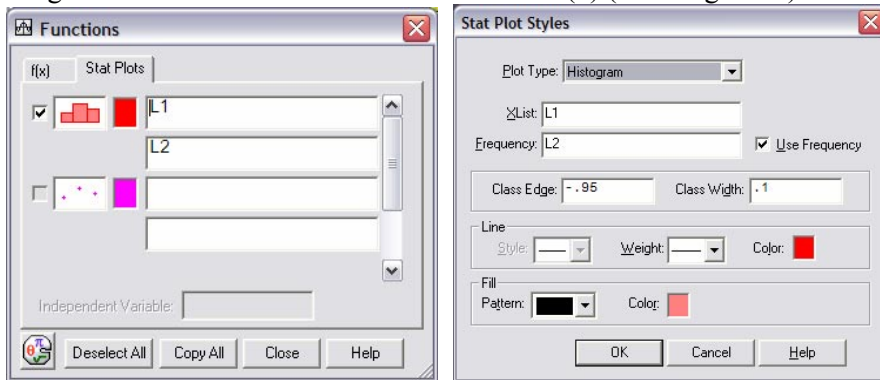
Öffnen Sie den Listeneditor, indem Sie auf  klicken.
 Klicken Sie nacheinander auf den Spaltenkopf der Listen L1, L2 und L3 und geben Sie in das erscheinende Eingabefenster für Formeln, die aus dem Bild ersichtliche Formel ein.



Nach dem Erstellen der Listen klicken Sie auf , um die Eingaben für die grafische Darstellung vorzunehmen.

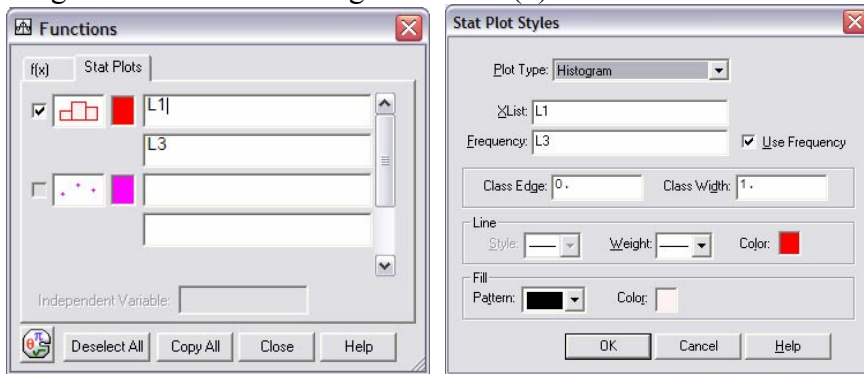


Eingaben für die Wahrscheinlichkeitsfunktion $f(x)$ (Stabdiagramm):

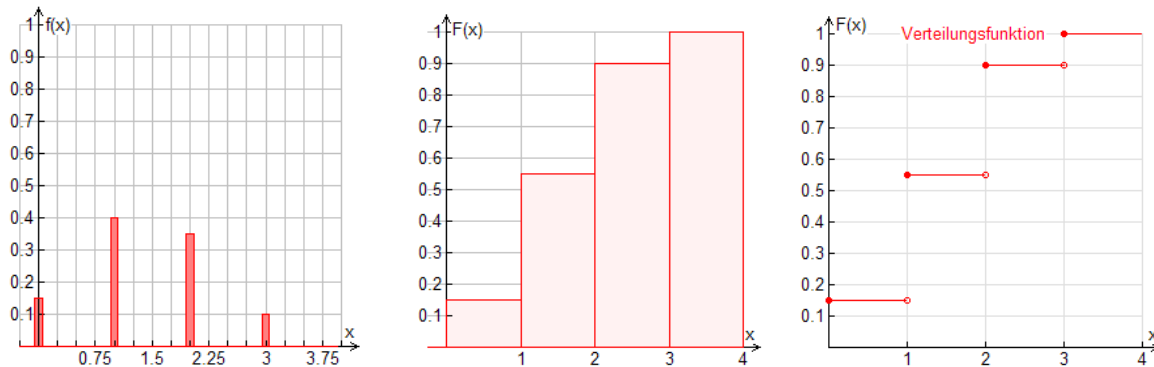


Fügen Sie die Grafik mit in Ihr Arbeitsblatt.

Klicken Sie nochmals auf , um auch die Verteilungsfunktion $F(x)$ zu zeichnen.
Eingaben für die Verteilungsfunktion $F(x)$:



Fügen Sie auch die zweite Grafik mit in Ihr Arbeitsblatt.



Anmerkung: Die exakte Darstellung der Verteilungsfunktion (Bild ganz rechts) ist sehr aufwändig zu erstellen. Wir begnügen uns in Folge mit der genäherten Darstellung als der Verteilungsfunktion als Histogramm (Bild in der Mitte).

Mit rufen Sie das Stat Calculation Tool auf.

Eingaben:

Calculation Type: One-Variable Statistics

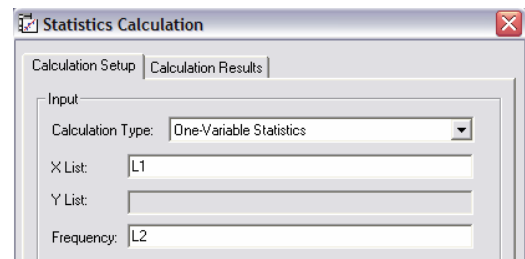
X List: L1

Frequency: L2

OK

Save Results

One-Variable Statistics
 $\bar{x} = 1.4$
 $\sigma_x = .864099$



Sie erhalten Erwartungswert und Standardabweichung berechnet.

Stichprobenanweisung

Ein Betrieb stellt monatlich eine sehr große Stückzahl eines Produktes A her und liefert dieses Produkt an einen Abnehmer. Der Produzent versichert, dass maximal $p = 5\%$ der von ihm gelieferten Produkte fehlerhaft sind.

Zur Sicherung der Qualität wird jede Lieferung mit Stichproben überprüft. Beim Abschluss des Liefervertrages einigt man sich darauf eine n-c-Stichprobenanweisung durchzuführen.

Bei jeder Lieferung wird eine Zufallsstichprobe vom Umfang $n = 100$ entnommen.

Die Zahl $c = 5$ ist die Annahmezahl.

Findet man bis zu 5 defekte Produkte in der untersuchten Lieferung, wird diese Lieferung trotzdem angenommen.

Findet man mehr als 5 defekte Produkte in der Lieferung, wird das Los (die gesamte Lieferung) abgelehnt.

Die Wahrscheinlichkeit dass eine Lieferung angenommen wird, heisst Annahmewahrscheinlichkeit P_a .

$$P_a = P(X \leq c) = F(c) = P(X = 0) + P(X = 1) + \dots + P(X = c)$$

p ist der Anteil der fehlerhaften Produkte in einem Los (Fehleranteil). (n_c_Stichprobenanweisung.tii)

Zur Berechnung der Annahmewahrscheinlichkeit eignet sich die Binomialverteilung. Für umfangreichere Berechnungen ist es oft sinnvoll eine eigene Funktion zu definieren.

```
yb(p) := binomCDF(n, p, c)           yb(p) = binomcdf(100, p, 5)
```

a) Wie groß ist die Annahmewahrscheinlichkeit bei einem Fehleranteil von $p = 5\%$?

$$P_a = P(X \leq 5) = \text{binomcdf}(100, .05, 5) = 0.616$$

```
binomcdf(n, 0.05, c) = .615999128   yb(.05) = .615999
```

Bei einem Fehleranteil von 5% werden ca. 61.6% der Lose angenommen.

b) Wie groß ist die Annahmewahrscheinlichkeit bei einem Fehleranteil von $p = 10\%$?

$$P_a = P(X \leq 5) = \text{binomcdf}(100, .1, 5) = 0.0576$$

```
binomcdf(100, 0.1, 5) = .057576886
```

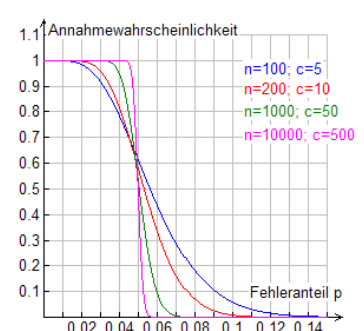
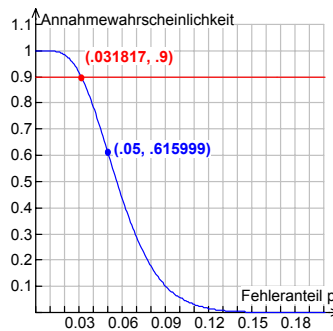
Bei einem Fehleranteil von 10% werden nur noch ca. 5.76% der Lose angenommen.

c) Wie hoch darf der Fehleranteil p maximal sein, wenn die Annahmewahrscheinlichkeit $P_a = 90\%$ sein soll. Bei TII ist die Berechnung mit solve(und nsolve(ist nicht zielführend. Man muss zusätzlich ein Intervall für p angeben!

```
nsolve(binomcdf(n, p, c) = 0.9, p)  EVAL ERROR: Probability must be in [0,1] interval.
nsolve(binomcdf(n, p, c) = 0.9, p) | p >= 0 and p <= 1 = .031817
```

Einfacher ist die grafische Berechnung des Schnittpunktes der Geraden $y_2(x)=0.9$ mit der Kennkurve in der Grafik.

- Eingabe TII: $y_1 = \text{binomCDF}(100, p, 5)$
 $y_2 = 0.9$
- Vergleiche: $y_1 = \text{binomCDF}(100, p, 5)$
 $y_2 = \text{binomCDF}(200, p, 10)$
 $y_3 = \text{binomCDF}(1000, p, 50)$
 $y_4 = \text{binomCDF}(10000, p, 500)$



Beispiel Überbuchung: Ein Ferienhotel hat 300 Zimmer. Der Direktor weiss, dass durchschnittlich 5% aller reservierten Zimmer nicht belegt werden. Aus diesem Grund wird das Hotel manchmal überbucht. Es werden mehr als 300 Zimmerbestellungen entgegengenommen.

a) Wie viele Buchungen dürfen höchstens angenommen werden, wenn mit einer Wahrscheinlichkeit von mindestens 95% nicht zu viele Gäste (= höchstens 300 Gäste/Zimmerbelegungen) eintreffen sollen. Dh. die Wahrscheinlichkeit, dass das Hotel überbucht ist, darf 5% nicht übersteigen.

X = Anzahl der tatsächlich eingetroffenen Gäste; $p = 0.95$

n Stichprobenumfang; Anzahl der Buchungen; gesucht

$p = 0.95$ Erfolgswahrscheinlichkeit; Gast kommt tatsächlich

$$m = n \cdot p \quad ; \quad s = \sqrt{n \cdot p \cdot (1 - p)}$$

$$\sigma = .217945 \cdot \sqrt{n} \quad ; \quad \mu = .95 \cdot n$$

Es soll gelten: $P(X > 300) \leq 0.05$

$$1 - P(X \leq 300) \leq 0.05$$

Die Wahrscheinlichkeit, dass **mehr als 300 Gäste** kommen soll **maximal 0.05** sein.

Die Berechnung der zulässigen Buchungen n erfolgt durch Probieren oder mit einer Tabelle.

Probieren: $n=310$: $1 - \text{binomCDF}(310, p, 300) = .050897707704 > 0.05$

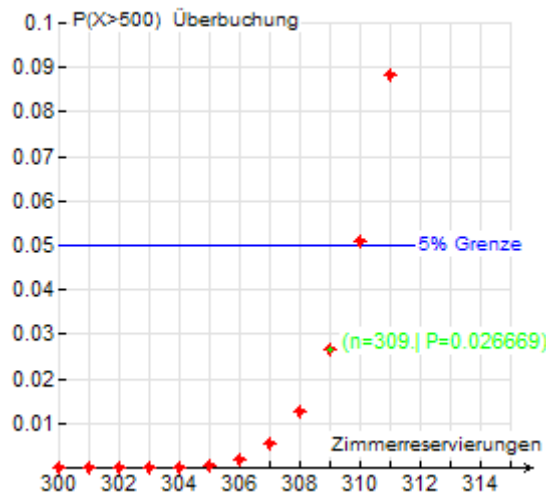
$n=309$: $1 - \text{binomCDF}(309, p, 300) = .026668975875 < 0.05$

Somit dürfen maximal **309 Gäste** gebucht werden!

$\text{seq}(x, x, 300, 350) \rightarrow L1$

$\text{seq}(1 - \text{binomCDF}(x, p, 300), x, 300, 350) \rightarrow L2$

L1	L2
300	0
301	2E-007
302	3E-006
303	3E-005
304	0.00014
305	0.00057
306	0.00188
307	0.00522
308	0.01256
309	0.02667
310	0.0509
311	0.08845
312	0.14154
313	0.21056
314	0.29364



In Liste L2 steht die Wahrscheinlichkeit, dass das Hotel **überbucht** ist.

Der Direktor darf maximal 309 Buchungen akzeptieren. Je weniger Buchungen der Direktor zulässt, umso sicherer ist, dass das Hotel nicht überbucht ist. Bei 300 akzeptierten Buchungen ist das Hotel sicher nicht überbucht. Ab 310 Buchungen ist das Hotel mit mehr als 5%-iger Wahrscheinlichkeit überbucht.

Durchschnittlich werden $\mu = n \cdot p = 309 \cdot 0.95 = 293.55 \approx 294$ Zimmer belegt sein.

www.t3oesterreich.at

Simulation Augenzahl eines Würfels

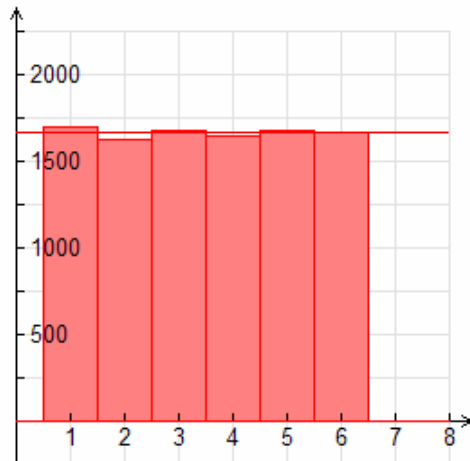
procedure

Augensumme von $n = 1$ Würfel bei $k := 10000$ Versuchsdurchführungen

`randseed(123) = "Done"` Initialisierung des Rechners

`randint(1, 6, k) → L1` Erzeugt eine Liste mit 10000 Zufallszahlen zwischen 1 und 6 (inkl.).

$$D := \frac{k}{6} = \frac{5000}{3}$$



One-Variable Statistics

$$\bar{x} = 3.4977$$

$$\Sigma x = 34977.$$

$$\Sigma x^2 = 151667.$$

$$S_x = 1.71263$$

$$\sigma_x = 1.71254$$

$$n = 10000.$$

$$\min X = 1.$$

$$Q1 = 2.$$

$$\text{Median} = 3.$$

$$Q3 = 5.$$

$$\max X = 6.$$

Es zeigt sich, dass sich die real ermittelten statistischen Kennzahlen nur wenig von den theoretischen Werten μ und σ abweichen.

Beispiel Simulation:

k-maliges Werfen von zwei Würfeln;

$X = X_1 + X_2 =$ Summe der Augenzahlen; X_i Augenzahl des i-ten Würfels.

Simulation mit TI

TI InterActive! hat gegenüber den Handhelds den Vorteil vom Speicherplatz (fast) nicht beschränkt zu sein.

Es lassen sich somit Simulationen mit einer sehr großen Zahl k von *Durchführungen* rechnen.

Die Dateien werden allerdings sehr groß, können aber durch „Zippen“ deutlich verkleinert werden.

Initialisieren Sie zunächst Ihren Rechner wieder mit **123**.

Die Vorgangsweise ist hier etwas komplizierter als bei der ersten Simulation und soll daher in Schritten erklärt werden. Die Formel für die eigentliche Berechnung ergibt sich erst im letzten Schritt.

- `randInt(1,6)` erzeugt eine ganzzahlige Zufallszahl zwischen 1 und 6
- `randInt(1,6,2)` erzeugt eine Liste aus zwei Zufallszahlen zwischen 1 und 6.
- `sumlist(randInt(1,6,2))` berechnet die Summe der Werte einer Liste mit zwei Zufallszahlen zwischen 1 und 6.
- `seq(sumlist(randInt(1,6,2)),X,1,12000)` berechnet eine Liste mit 12000 Summen zweier Zufallszahlen (Augensummen von 2 Würfeln). Diese Liste soll wieder in L1 gespeichert werden.

$$E(X) = \bar{x} \approx 7 = \mu = E(X_1) + E(X_2) = 3.5 + 3.5 = 2 \cdot 3.5$$

$$\sigma = 2.249 \approx \sqrt{2} \cdot 1.7078 = 2.4$$

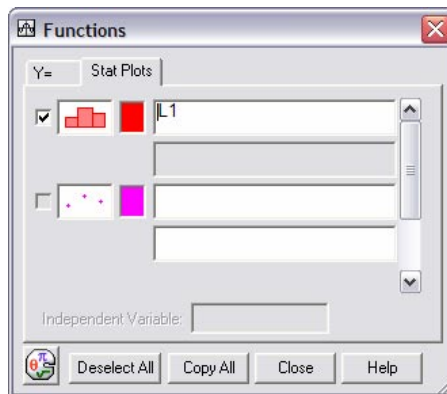
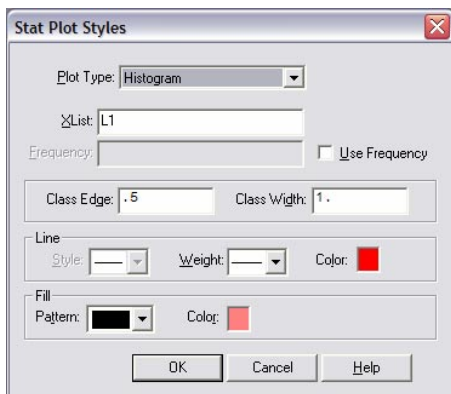
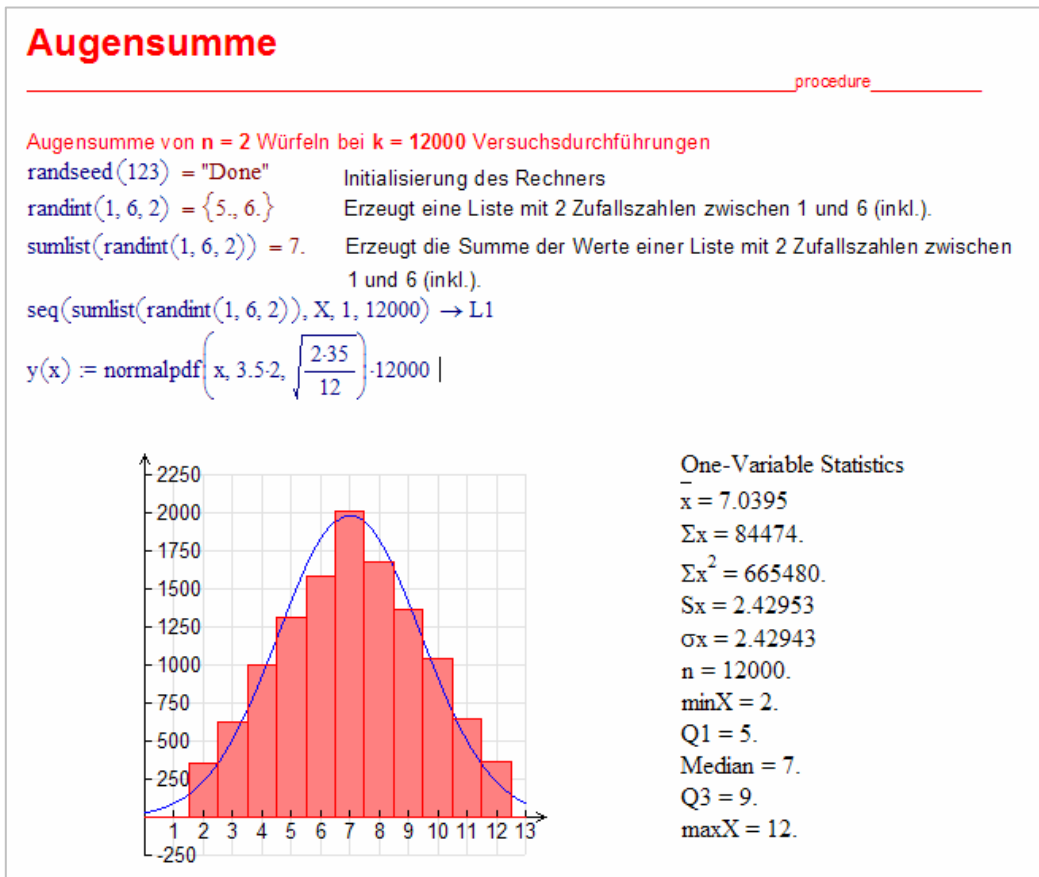
(Anmerkung: $\sigma = s_{n-1} = s_x$)

$$V(X) = 2.25^2 = 5.06 \approx 2 \cdot \sigma^2 = 2 \cdot 35/12 = 5.8$$

Mit der Funktion $Y1 = \text{normalpdf}(X, n \cdot 3.5, \sqrt{(n \cdot 35/12)}) \cdot k$ kann die Güte der Approximation grafisch überprüft werden.

Eingabe: $Y1 = \text{normalpdf}(X, 2 \cdot 3.5, \sqrt{(2 \cdot 35/12)}) \cdot 12000$

(Simulation2Wuerfel.tii; Simulation5Wuerfel.tii;)



Beispiel:

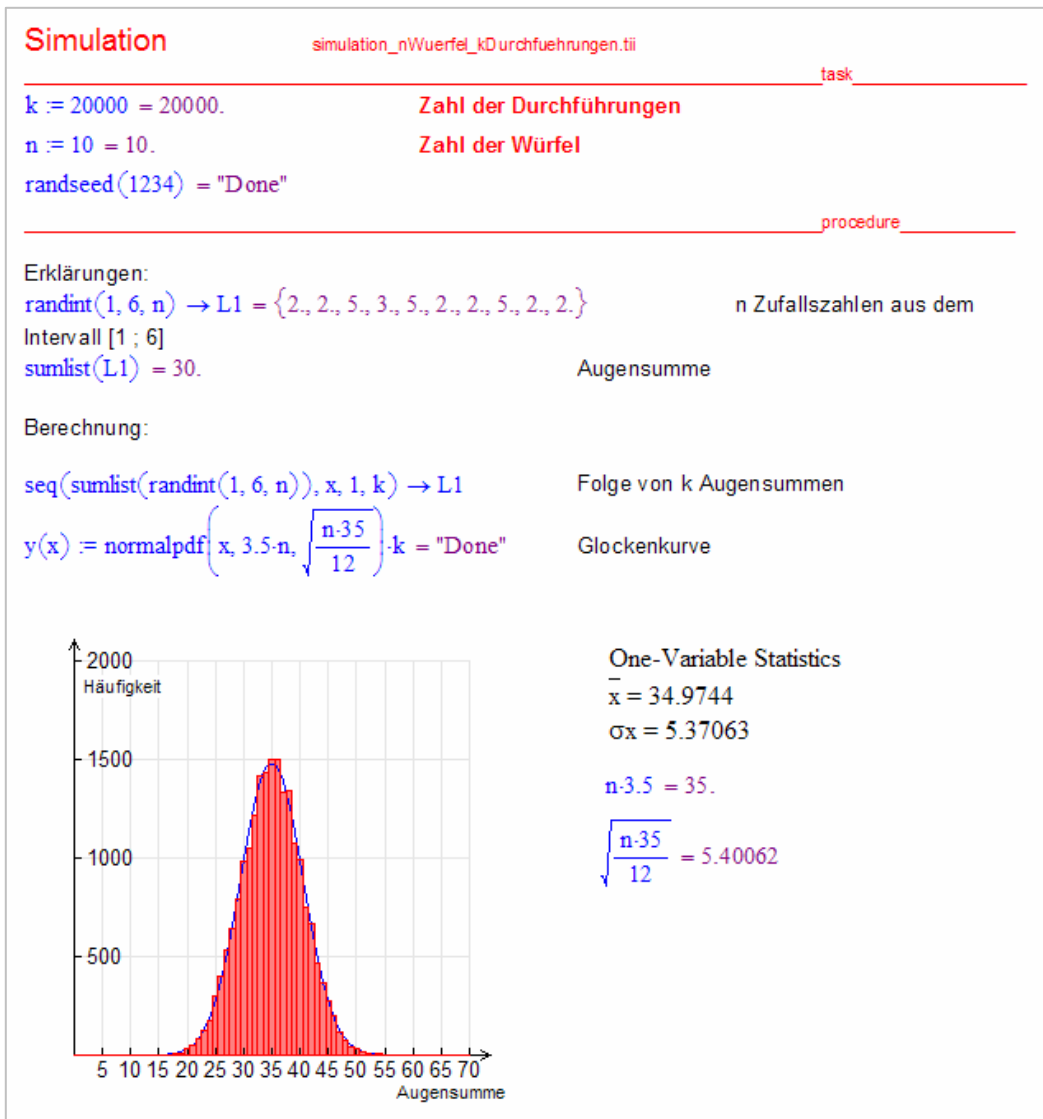
k-maliges Werfen von 10 Würfeln;

$X = X_1 + X_2 + \dots + X_{10}$ = Summe der Augenzahlen ; X_i Augenzahl des i-ten Würfels.

Vermutung für den Erwartungswert $E(X)$ der Augensumme und die Varianz $V(X)$ der Augensumme:

$$E(X) = 10 \cdot 3.5 = 35; \quad V(X) = 10 \cdot \frac{35}{12} = 29.167$$

(simulation_nWuerfel_kDurchfuehrungen.tii)



Aufgrund der vorliegenden Ergebnisse *vermuten* wir für $X = X_1 + X_2 + X_3 + \dots + X_n$

Erwartungswert: $E(X) = E(X_1) + E(X_2) + E(X_3) + \dots + E(X_n)$

Varianz: $V(X) = V(X_1) + V(X_2) + V(X_3) + \dots + V(X_n)$

Anmerkung: X_i und X_j sind paarweise unabhängig

Beispiel Simulation:

Die n -gewichteten Varianz der Stichprobe ist **nicht** gleich der Varianz der Grundgesamtheit!
 Dies gilt aber für die $(n-1)$ -gewichtete Varianz!
 Mit Hilfe einer Simulation soll dieser Sachverhalt veranschaulicht werden.

S_{n-1}^2 ist ein **erwartungsgerechter Schätzer** der Varianz der Grundgesamtheit.

Die unbekannte Varianz σ^2 der Zufallsvariablen X (Grundgesamtheit) ist durch $\sigma^2 \approx S_{n-1}^2$ zu schätzen.

Einen Beweis des Satzes finden Sie in Kreyszig, Seite 392.
 (simulation_sn-1.tii)

Simulation $\sigma = s_{n-1}$

Simulation_sn-1.tii

procedure

```
randseed(1234) = "Done"
```

```
n := 5 = 5    μ := 50 = 50
```

```
σ := 10 = 10  Dieser Wert soll geschätzt werden:  $\sigma^2 = 100$ 
```

1. Schritt: 5 normalverteilte Zufallszahlen:

```
randnorm(μ, σ, n) → L1 = {56.9021, 54.9097, 45.4352, 50.9069, 41.0626}
```

2. Schritt:

Berechnung der $(n-1)$ -Varianz der Zufallszahlen für **eine** Durchführung der Stichprobe vom Umfang n :

```
variance(L1) = 43.2897
```

Berechnung der n -Varianz der Zufallszahlen für **eine** Durchführung der Stichprobe vom Umfang n :

```
 $\frac{n-1}{n} \cdot \text{variance}(L1) = 34.6317$ 
```

3. Schritt: Folge von 1000 Durchführungen vom Umfang $n = 5$

```
seq(variance(randnorm(μ, σ, n)), x, 1, 1000) → L1
```

4. Schritt:

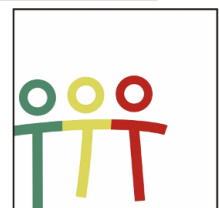
Mittelwert der berechneten $(n-1)$ - Varianzen:

```
mean(L1) = 99.153
```

Mittelwert der berechneten n -Varianzen:

```
mean( $\left(\frac{n-1}{n} \cdot L1\right)$ ) = 79.3224
```

Welcher der beiden Werte schätzt $\sigma^2 = 100$ besser?

T³ ÖSTERREICH

3 Konfidenzintervalle; Vertrauensbereiche für den Anteilswert π

Ein Konfidenzintervall (Vertrauensbereich) mit dem **Konfidenzniveau c** ist ein Bereich, in dem ein unbekannter Parameter mit der Wahrscheinlichkeit c liegt.

Mit der **Irrtumswahrscheinlichkeit $\alpha = 1 - c$** liegt der Parameter nicht im berechneten Intervall.

Meist verwendet man **$\alpha = 5\%$** ($c = 95\%$; 95%-Konfidenzintervall) oder **$\alpha = 1\%$** ($c = 99\%$; 99%-Konfidenzintervall).

Beispiel:

Ein 95%-Konfidenzintervall überdeckt den gesuchten Parameter der Grundgesamtheit in ca. 95 von 100 berechneten Intervallen.

Ca. 5 der 100 Intervalle überdecken den gesuchten Parameter der Grundgesamtheit nicht.

Wir betrachten in diesem Kapitel nur zweiseitige Konfidenzintervalle für Anteilswerte.

Beispiel:

Eine Telefonumfrage unter 300 Wahlberechtigten ergab, dass bei der nächsten Wahl $x = 140$ von den fragten Personen den Kandidaten der Partei A wählen würden. Der Rest der Stimmen verteilt sich auf andere Parteien. Berechnen Sie ein 95% Konfidenzintervall für den **unbekannten** Wähleranteil π der Partei A unter allen Wahlberechtigten.

Die Anzahl $X =$ **Wähler der Partei A** in einer Stichprobe vom Umfang n ist binomialverteilt $B(n; \pi)$ mit dem **bekanntem** Parameter n und einem **unbekanntem** $\mu = n \cdot \pi$.

Der Anteilswert π der Grundgesamtheit ist **unbekannt**.

$x = 140$ beabsichtigte Wähler von A in Stichprobe

$n = 300$ Umfang der Stichprobe

$c = 95\%$ Niveau des Konfidenzintervalls

$\hat{p} = \frac{140}{300} \approx 0.4667$ Wähleranteil (bekannt) von A in der Stichprobe

$$\hat{p} - z \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \leq \pi \leq \hat{p} + z \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \quad \text{mit } z = \Phi^{-1}\left(\frac{c+1}{2}\right) \quad \text{für } 0.3 < \hat{p} < 0.7$$

$$[\hat{p} - e; \hat{p} + e] \quad \text{mit } e = z \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \quad e \text{ heißt } \mathbf{Fehlertoleranz} \text{ oder } \mathbf{Schwankungsbreite}$$

Diese Näherungsformel findet für $n > 30$ und $n \cdot \hat{p} \cdot (1 - \hat{p}) > 9$ findet in der Praxis sehr häufig verwendet.


Alle TI-Produkte rechnen nur mit dieser Näherungsformel!

Fortsetzung des letzten Beispiels

Berechnung des Konfidenzintervalls

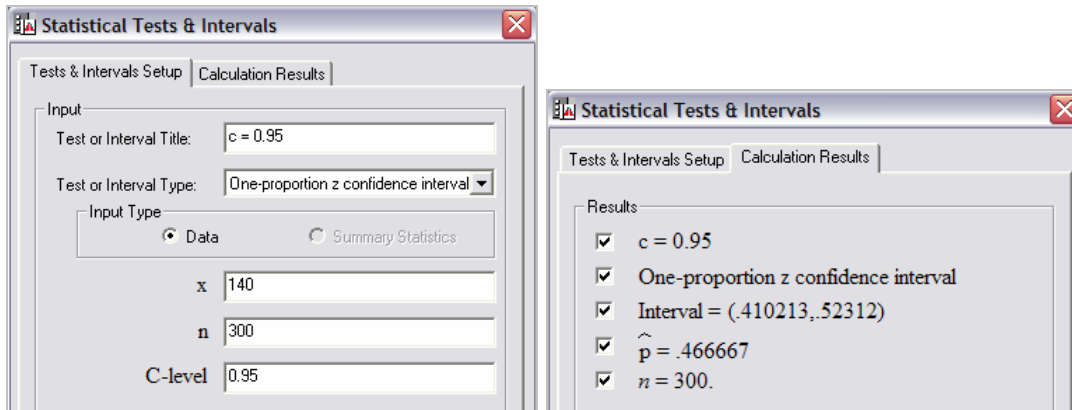
$$\left[\frac{140}{300} - 1.96 \cdot \sqrt{\frac{\frac{140}{300} \cdot \left(1 - \frac{140}{300}\right)}{300}}; \frac{140}{300} + 1.96 \cdot \sqrt{\frac{\frac{140}{300} \cdot \left(1 - \frac{140}{300}\right)}{300}} \right] = [0.410213; 0.52312]$$

Interpretation: Mit einer Wahrscheinlichkeit von 95% liegt der Wähleranteil des Kandidaten der Partei A zwischen 41% und 52.3% der abgegebenen Stimmen.

Bei Verwendung von **TII** klicken Sie auf  und öffnen damit ein Fenster zur Berechnung von Konfidenzintervallen.

Tragen Sie die gegebenen Werte in die dafür vorgesehenen Eingabefenster ein und klicken Sie auf

Calculate. Mit **Save Results** fügen Sie die Ergebnisse in Ihr Arbeitsblatt ein.



Wie ist das Konfidenzintervall nun zu interpretieren?

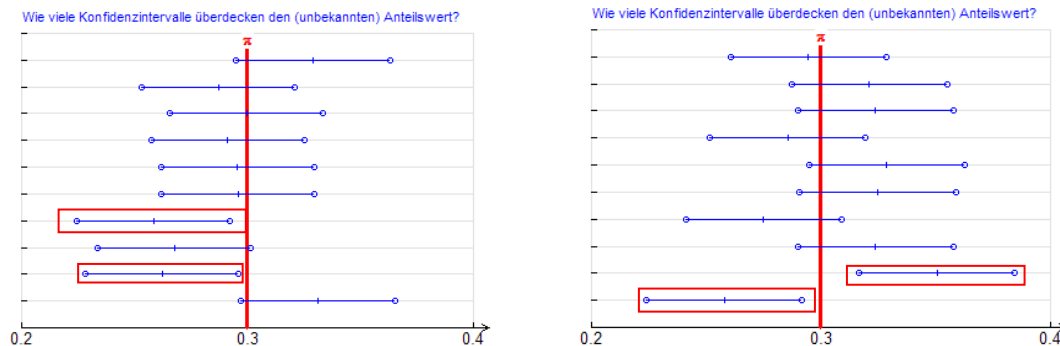
Nehmen wir an, wir ziehen 100 Stichproben zu je 300 Personen und erheben den Anteilswert der Wähler von Kandidat A in diesen 100 Stichproben sowie die Standardabweichung und die Grenzen des Konfidenzintervalls zum Signifikanzniveau 95 %.

Dann können wir erwarten, dass 95 der 100 Konfidenzintervalle den Anteilswert der Grundgesamtheit enthalten, aber 5 Konfidenzintervalle enthalten den Anteilswert der Grundgesamtheit nicht.

Beachten Sie: Ein einzelnes Konfidenzintervall enthält entweder den Grundgesamtheitsparameter π oder es enthält ihn nicht. Ob es den Grundgesamtheitsparameter enthält, können wir nur mit einer Sicherheit behaupten, die durch das Konfidenzniveau c (meistens 95 %) angegeben ist.

Die folgende Grafik zeigt 10 Konfidenzintervalle von 10 verschiedenen Stichproben aus einer Grundgesamtheit mit dem Anteilswert π zum Konfidenzniveau 80 %.

8 der 10 Konfidenzintervalle enthalten **im Schnitt** den Grundgesamtheitsparameter π , zwei aber nicht! ([konfidenzsimulatAnteilswert.tii](#))



Grafische Darstellung von Konfidenzintervallen durch „Konfidenzellipsen“

Für das Konfidenzintervall des Anteilswertes π einer Grundgesamtheit gilt

$$\hat{p} - e \leq \pi \leq \hat{p} + e \quad \text{mit } \hat{p} = \frac{x}{n} \text{ und } e = z \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \quad \text{und } z = \Phi^{-1}\left(\frac{c + 1}{2}\right)$$

Der Anteilswert der Grundgesamtheit π liegt also mit Wahrscheinlichkeit c zwischen einer oberen und einer unteren Grenze des Konfidenzintervalls. Zur grafischen Veranschaulichung von Konfidenzintervallen werden zuerst die Gleichungen der begrenzenden Funktionen $y_u(n, \hat{p})$ und $y_o(n, \hat{p})$ - das sind gleichzeitig auch die Grenzen des Konfidenzintervalls - definiert.

Ausserdem wird die Funktion für die Schwankungsbreite $e(n, \hat{p})$ definiert.

Schwankungsbreite e :
$$e(n, \hat{p}) = z \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \quad \text{mit } z = \Phi^{-1}\left(\frac{c + 1}{2}\right)$$

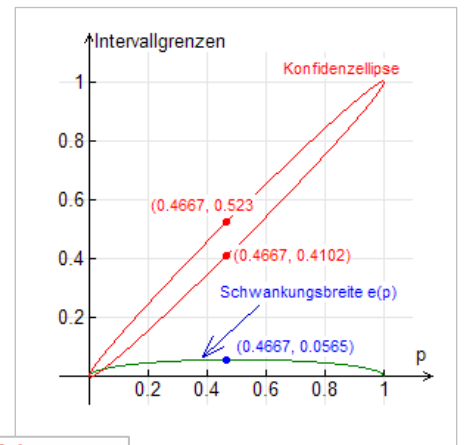
Untere begrenzende Funktion: $y_u(n, \hat{p}) = \hat{p} - e(n, \hat{p})$

Obere begrenzende Funktion: $y_o(n, \hat{p}) = \hat{p} + e(n, \hat{p})$

Diese beiden Funktionen $y_u(n, \hat{p})$ und $y_o(n, \hat{p})$ werden dann für konkrete Werte von n und \hat{p} als unabhängige Variable in einem passenden Koordinatensystem dargestellt.

Die Grafik rechts zeigt die Konfidenzellipse für $n = 300$; $x = 140$; $c = 0,95$.

In der Grafik geben die y-Koordinaten zweier Punkte auf der Konfidenzellipse die Grenzen des Konfidenzintervalls für den Wert $\hat{p} = 0,4667$ an. (Anmerkung: Aus technischen Gründen wird in der Grafik p anstelle von \hat{p} verwendet.) (Konfidenzellipse.tii, konfidenz.tii)

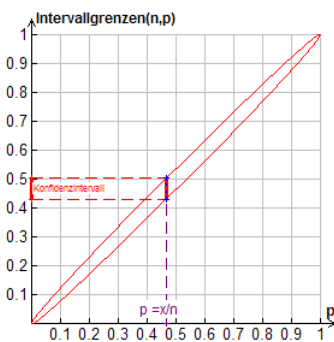


$z := \text{invNorm}\left(\frac{c+1}{2}\right) = \text{invnorm}\left(\frac{c+1}{2}\right)$	$x := 140 = 140.$ $n := 300 = 300.$ $c := 0.95 = .95$	Zahl der Erfolge Umfang d. Stichprobe Konfidenzniveau
$yU(n, p) := p - z \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$	$\hat{p} = p := \frac{x}{n} = .466667$	
$yO(n, p) := p + z \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$	$n \cdot p \cdot (1-p) = 74.6667$ Kontrolle > 9 $yU(n, p) = .410213$ $yO(n, p) = .52312$	

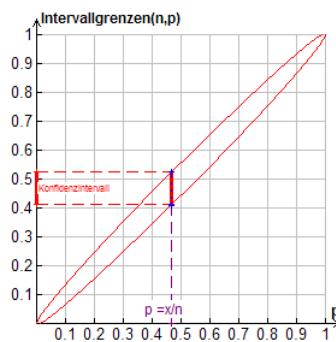
Konfidenzintervall: [0,410; 0,523]

Die folgenden Bilder zeigen „Konfidenzellipsen“ für verschiedene Konfidenzniveaus. Es zeigt sich, dass die „Dicke“ der Ellipsen mit dem Konfidenzniveau c steigt.

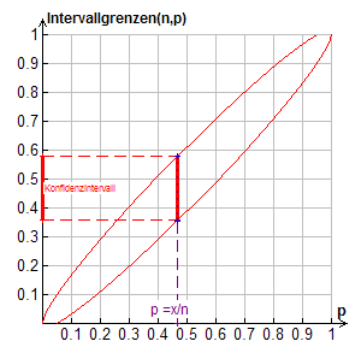
$n = 300$; $x = 140$; $c = 0,80$



$n = 300$; $x = 140$; $c = 0,95$

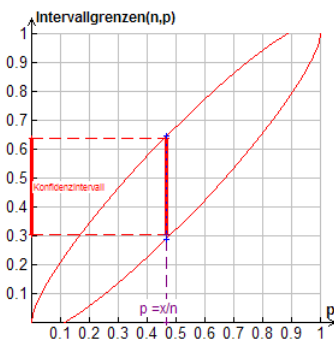


$n = 300$; $x = 140$; $c = 0,9999$

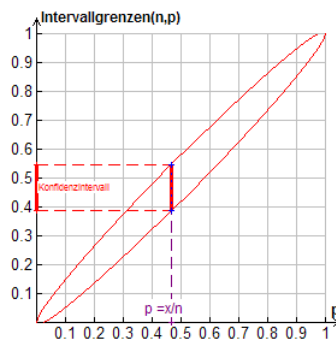


Eine Vergrößerung des Stichprobenumfangs n führt zu einer Verringerung der Breite eines Konfidenzintervalls,

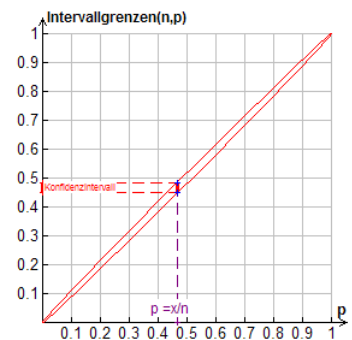
$n = 30$; $x = 14$; $c = 0,95$



$n = 150$; $x = 70$; $c = 0,95$



$n = 3000$; $x = 1400$; $c = 0,95$



Beispiel: Meinungsforschung (Schwankungsbreite.tii/pdf)

>> UMFRAGE

Auftraggeber: Bezirksblätter Burgenland

Ausführende Gesellschaft: GMK Gesellschaft für Marketing und Kommunikation, Graz

Zielgruppe: Bevölkerung des Burgenlandes ab 18 Jahren

Abfragezeitraum: 24. - 25. Februar 2005

Sample und Instrument: 402 Telefoninterviews

Maximale Schwankungsbreite: Gesamtergebnis ± 5%

Anmerkung: ± Werte im Vergleich zum letzten Politbarometer im März 2003.

Politbarometer Burgenland

Bekanntheitsgrad

Im Bezirksblatt vom März 2005 ist das Ergebnis einer Telefonumfrage zum **Bekanntheitsgrad** einiger burgenländischer Landespolitiker abgedruckt. Mit den Umfrageergebnissen ist auch der Stichprobenumfang mit $n = 402$ sowie die maximale Schwankungsbreite mit $\pm 5\%$ angegeben.

Berechnen Sie die gegebene Schwankungsbreite für die einzelnen Politiker. $c = 0.95$

Was ist unter dieser maximalen Schwankungsbreite zu verstehen? (Schwankungsbreite.tii)

Berechnungsformel:
$$e = z \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} = 1.96 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{402}} \quad z = \text{invnorm}(1.95/2) = 1.96$$

Niessl:
$$e = 1.96 \cdot \sqrt{\frac{1 \cdot (1 - 1)}{402}} = 0 \quad \pm 0\%; \quad \text{Jeder kennt Niessl!}$$

Steindl:
$$e = 1.96 \cdot \sqrt{\frac{0.94 \cdot (1 - 0.94)}{402}} = 0.023 \quad \pm 2.3\% \quad [91.7\% ; 96.3\%]$$

Illedits:
$$e = 1.96 \cdot \sqrt{\frac{0.58 \cdot (1 - 0.58)}{402}} = 0.0482 \quad \pm 4.82\% \quad [53.2\% ; 62.8\%]$$

Kölly:
$$e = 1.96 \cdot \sqrt{\frac{0.47 \cdot (1 - 0.47)}{402}} = 0.0488 \quad \pm 4.88\% \approx 5\% \quad [42.1\% ; 51.9\%]$$

Die Berechnungen zeigen, dass die Schwankungsbreite sehr stark vom Anteilswert $p (= \hat{p})$ der Stichprobe abhängt.

Hr. Niessl hat Schwankungsbreite 0, während das Konfidenzintervall von Hr. Kölly eine Breite von fast 10% hat.

Man kann allgemein zeigen, dass die maximale Schwankungsbreite bei $p = 0.5$ auftritt.

Aus der Grafik erkennt man, dass beispielsweise Umfragen mit $n = 100$ eine Schwankungsbreite von fast 10% (Breite des Konfidenzintervalls $\approx 20\%$) haben und somit unbrauchbar sind.

Sinnvoll sind Umfragen somit erst ab $n > 300$.

$$z := \text{invnorm}\left(\frac{1.95}{2}\right) = 1.95996$$

$$e(n, p) := z \cdot \sqrt{\frac{p \cdot (1 - p)}{n}} = \text{"Done"}$$

Functions

Y= Stat Plots

$y_1(p) := e(100, p)$

$y_2(p) := e(300, p)$

$y_3(p) := e(402, p)$

$y_4(p) := e(1000, p)$

Independent Variable: p

Deselect All Copy All Close Help

