Übung 3

-

Statistische Grundbegriffe und Ablauf der statistischen Untersuchung

Musterlösung

Aktuelle Version: 14. Juli 2022

Hinweise:

- Übungen sind mit Vorteil alleine zu lösen.
- Benutzen Sie die Musterlösungen nur zur Korrektur.
- Die Übungen sind wichtige Vorbereitungen für die Prüfung. Lösen sie die Übungen sorgfältig und stellen Sie die Lösungswege übersichtlich dar.
- (Ergänzte) Vorlesungsunterlagen und Fachbücher helfen beim Lösen von Übungen und bringen gleichzeitig eine erweiterte Ansicht auf die Problemstellung.
- Wenn Sie die Übungen nicht verstehen, fragen Sie!

Übung 1. *Fragen*

1. Erklären Sie den Unterschied zwischen einfacher und kumulierter Häufigkeitsverteilung.

   Die *einfache Häufigkeitsverteilung* zeigt die Häufigkeit (Anzahl) der einzelnen Werte, die in eine Klasse fallen. Bezieht man die Anzahl der Werte auf die Klassenweite, so kommt man zur Dichte. Das ist bereits eine Vorüberlegung zu der später in der Vorlesung eingeführten Dichtefunktion $f(x)$.

   Die *kumulierte Häufigkeitsverteilung* ist die schrittweise Summenbildung der einfachen Häufigkeitsverteilung. Der letzte Wert der Summenbildung ist damit die Anzahl aller erfassten Werte. Später werden wir hierfür auch den Begriff der Verteilungsfunktion $F(x)$ einführen.

   Überlegung: Betrachten wir die Dichtefunktion $f(x)$ und machen die Klassenintervalle x infinitesimal klein: $\Delta x \to dx$, dann geht die Summenbildung über in das Integral $F(x) = \int f(x)dx$ und kann als die Fläche unter $f(x)$ betrachtet werden.

2. Wann ist es erforderlich, eine klassifizierte Häufigkeitsverteilung zu erstellen und welcher Zielkonflikt ist bei der Klassenbildung zu lösen?

   Eine klassifizierte Häufigkeitsverteilung ist dann zu erstellen, wenn die Anzahl der zu erfassenden Werte zu gross wird, um eine übersichtliche Darstellung zu ermöglichen.

   Das Problem ist immer die Definition der Klassenweite. Zu gross oder zu klein, einheitlich oder nicht. Der Betrachter der klassifizierten Häufigkeitsverteilung geht wohl immer davon aus, dass die Verteilung innerhalb einer Klasse der Gleichverteilung entspricht. Das dürfte jedoch nicht immer der Falle sein.

3. Erklären Sie den Unterschied zwischen Primär-und Sekundärstatistik! Worin liegen jeweils die Vor- und Nachteile?

   Bei der *Primärstatistik* werden neue Daten für den Zweck der statistischen Untersuchung erhoben. Der Vorteil ist, dass die gesammelten Daten genau auf die Untersuchungsfrage ausgerichtet sind. Der Nachteil liegt darin, dass die Erhebung im Vergleich zur Sekundärstatistik teurer und aufwändiger ist.

   Bei der *Sekundärstatistik* werden bereits vorhandene Daten für eine neue statistische Untersuchung verwendet. Der Vorteil ist, dass die Datenerhebung kostengünstiger ist. Der Nachteil liegt darin, dass die Erhebung nicht auf die Fragestellung ausgerichtet worden ist. Bereits vorhandene Daten sind zudem weniger aktuell als neu erhobene Daten.

4. Erklären sie die Begriffe Merkmal, Merkmalswert, Merkmalsträger und Grundgesamtheit am Beispiel einer statistischen Erhebung der Durchschnittsgrösse einer Schulklasse.

Begriff	Beschreibung	Beispiel
Merkmal	zu erhebende Eigenschaft	Grösse
Merkmalsträger	Träger der Merkmale	Schüler
Merkmalswert	konkrete Ausprägung des Merkmals bei einem Träger	z.B. $1.65m$
Grundgesamtheit	Menge aller Merkmalsträger	alle Schüler der Klasse

5. Wofür werden Skalenniveaus verwendet? Welche Operationen sind mit ihnen möglich?

   Das Skalenniveau gibt Auskunft über die Qualität der Daten eines Merkmals, d.h. inwiefern sich das Merkmal in Zahlen darstellen lässt und inwieweit sinnvolle logische und mathematische Operationen angewendet werden können.

Nominalskala	$=\neq$
Ordinalskala	$=\neq<>$
Intervallskala	$=\neq<>+-$
Verhältnisskala	$=\neq<>+-*/$

6. Wie können Sie ein Histogramm manipulieren"?

   Ein Histogramm ist in erster Linie eine Visualisierung, weshalb Sie über eine unkorrekte Darstellung falsche Eindrücke vermitteln können:

   (a) Über die Wahl der Achsenmassstäbe können Verteilungen vorgetäuscht werden, z.B. eine schmalgipflige Verteilung durch Stauchung der x-Achse und Streckung der y-Achse.

   (b) Durch das Auslassen eines Teiles der y-Achse können grössere Differenzen vorgetäuscht werden, als tatsächlich vorhanden sind.

   (c) Durch die gleichförmige Darstellung von ungleichen Klassengrössen können Verteilungen vorgetäuscht werden, z.B. kann durch die Stauchung von grossen Randklassen eine schmalgpipflige Verteilung vorgetäuscht oder Rechts-/Linkssteilheit verborgen werden.

   Sie können auch über die Wahl der Klassen falsche Eindrücke vermitteln:

   (a) Durch die Wahl der Anzahl Klassen können Verteilungsformen vorgetäuscht werden, z.B. können bimodale Verteilungen durch die Wahl einer kleinen Anzahl von Klassen verdeckt werden.

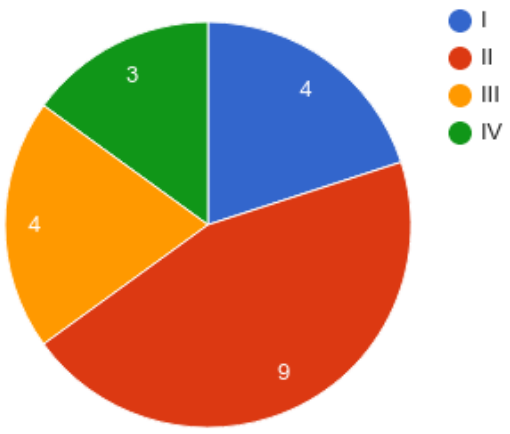
   (b) Durch die Wahl der Klassengrenzen- und grössen können ebenfalls Verteilungen unkorrekt dargestellt werden. Die Grenzen bestimmen, welche Datenpunkte zu einem neuen Datenpunkt zusammengefasst werden.

Übung 2. *Häufigkeiten*
Gegeben ist die folgende Tabelle der Firma Gut:

Nr.	Name Vorname	Familienstand	Zahl der Kinder	Tarifgruppe
1	Amberger, Heim	ledig	0	II
2	Bauer, Regine	verheiratet	2	I
3	Bertram, Günther	verheiratet	1	II
4	Dünnes, Rita	geschieden	0	I
5	Engel, Erika	ledig	1	II
6	Jahauf, Ernst	verheiratet	1	III
7	Prisch, Anton	verwitwet	3	II
8	Gillhuber, Erwin	verheiratet	0	III
9	Hell, Marion	geschieden	0	II
10	Jahn, Josef	ledig	2	II
11	Kaps, Wolfgang	verheiratet	0	III
12	Fritzen, Ernst	verwitwet	4	II
13	Lechner, Ernst	verheiratet	0	II
14	Kavlaier, Waltraud	ledig	1	I
15	Vafayer, Elisabeth	ledig	1	IV
16	Pagler, Fritz	ledig	2	IV
17	Polzer, Herrmaanr	verheiratet	3	III
18	Feiser, Ciabriele	geschieden	2	II
19	Schnidt, Heinz	geschieden	1	IV
20	Wenisch, Willy	verheiratet	0	I

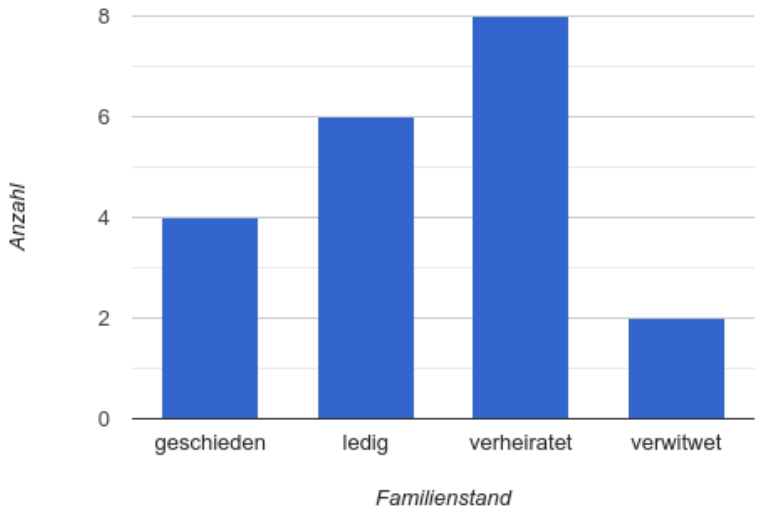
1. Berechnen Sie die absolute einfache, die absolute kumulierte, die relative einfache sowie die relative kumulierte Häufigkeit der Tarifgruppe. Erstellen Sie ein Kreisdiagramm der einfachen Häufigkeiten.

Tarifgruppe	h_i	H_i	f_i	F_i
I	4	4	0.2	0.2
II	9	13	0.45	0.65
III	4	17	0.2	0.85
IV	3	20	0.15	1.0



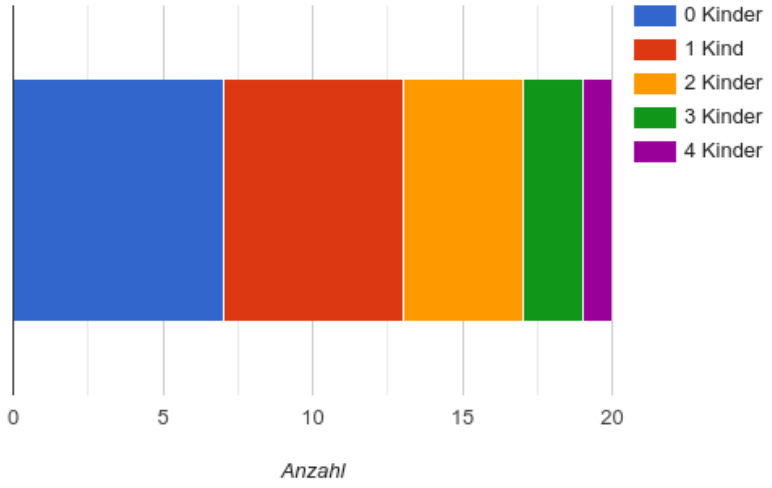
2. Berechnen Sie die absolute einfache, die absolute kumulierte, die relative einfache sowie die relative kumulierte Häufigkeit des Familienstands (geschieden, ledig, verheiratet, verwitwet). Erstellen Sie ein Stab-/Säulendiagramm der absoluten einfachen Häufigkeit.

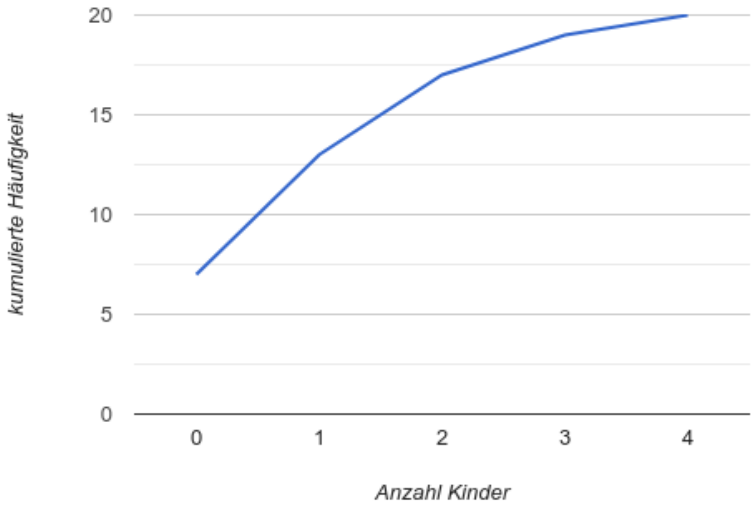
Familienstand	h_i	H_i	f_i	F_i
geschieden	4	4	0.2	0.2
ledig	6	10	0.3	0.5
verheiratet	8	18	0.4	0.9
verwitwet	2	20	0.1	1.0



3. Berechnen Sie die absolute einfache, die absolute kumulierte, die relative einfache sowie die relative kumulierte Häufigkeit der Anzahl Kinder. Erstellen Sie ein Rechteckdiagramm der absoluten einfachen Häufigkeit. Erstellen Sie ein Polygonzug der absoluten kumulierten Häufigkeit.

Anzahl Kinder	h_i	H_i	f_i	F_i
0	7	7	0.35	0.35
1	6	13	0.3	0.65
2	4	17	0.2	0.85
3	2	19	0.1	0.95
4	1	20	0.05	1.0





Übung 3. *Klassen*
Sie haben folgende Urliste mit Werten:

131.8	106.7	116.4	84.3	118.5	93.4	65.3	113.8	140.3
119.2	129.9	75.7	105.4	123.4	64.9	80.7	124.2	110.9
86.7	112.7	96.7	110.2	135.2	134.7	146.5	144.8	113.4
128.6	142.0	106.0	98.0	148.2	106.2	112.7	70.0	73.9
78.8	103.4	112.9	126.6	119.9	62.6	116.6	84.6	101.0
68.1	95.9	119.7	122.0	127.3	109.3	95.1	103.1	92.4
103.0	90.2	136.1	109.6	99.2	76.1	93.9	81.5	100.4
114.3	125.5	121.0	137.0	107.7	69.0	79.0	111.7	98.8
124.3	84.9	108.1	128.5	87.9	102.4	103.7	131.7	139.4
108.0	109.4	97.8	112.2	75.6	143.1	72.4	120.6	95.2

Erstellen Sie ein Balkendiagramm mit den Klassenhäufigkeiten. Verwenden Sie für die Berechnung der Anzahl Klassen die Faustregel von Sturges (1926): $m \approx 1 + 3.32 \cdot log(n)$ und eine ganzzahlige Klassenbreite.

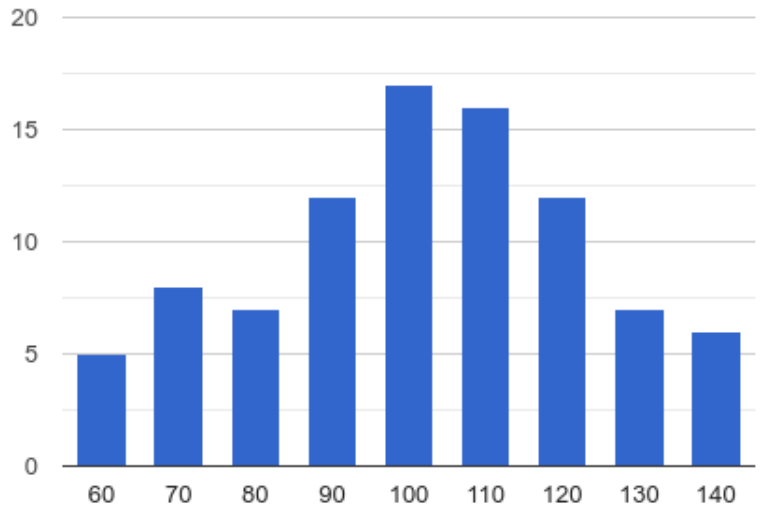
Wir berechnen die Anzahl Klassen m und die Klassenbreite Kb mit der Fausregel von Sturges:

$$m \approx 1 + 3.32 \cdot log(90) \approx 8$$

$$Kb = \frac{x_{max} - x_{min}}{m} = \frac{148.2 - 62.6}{8} = 10.7 \approx 10$$

Durch die Verwendung einer zusätzlichen Klasse können wir anschauliche Klassen in Zehnerschritten machen:

i	x	h_i
1	60.0-69.9	5
2	70.0-79.9	8
3	80.0-89.9	7
4	90.0-99.9	12
5	100.0-109.9	17
6	110.0-119.9	16
7	120.0-129.9	12
8	130.0-139.9	7
9	140.0-140.9	6



Übung 4. *Klassenhäufigkeiten*
Die Brenndauer von 200 Glühbirnen ist folgendermassen verteilt:

Brenndauer in Stunden		Anzahl Glühbirnen
von	bis	
0	4000	12
4000	6000	28
6000	7000	44
7000	8000	68
8000	9000	30
9000	10000	18

1. Bestimmen Sie die einfachen, relativen und die kumulierten Klassenhäufigkeiten! Interpretieren Sie die Werte h_2, f_4, H_3 und F_5!

i	x_i^u	x_i^o	h_i	H_i	f_i	F_i
1	0	4000	12	12	0.06	0.06
2	4000	6000	28	40	0.14	0.20
3	6000	7000	44	84	0.22	0.42
4	7000	8000	68	152	0.34	0.76
5	8000	9000	30	182	0.15	0.91
6	9000	10000	18	200	0.09	1.00

Interpretation:

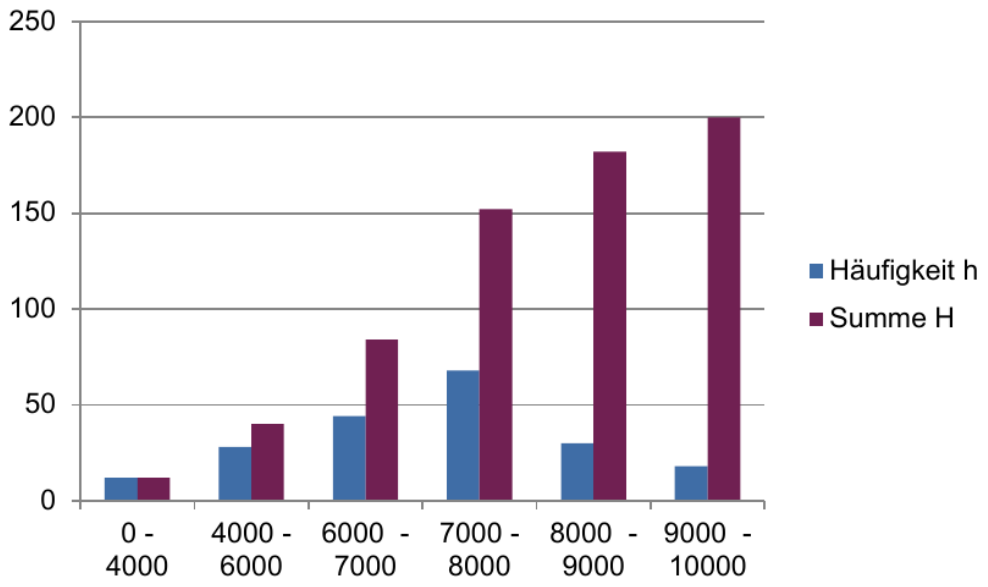
(a) $h_2 = 28$: 28 Glühbirnen haben eine Brenndauer zwischen 4000 und 6000 Stunden.

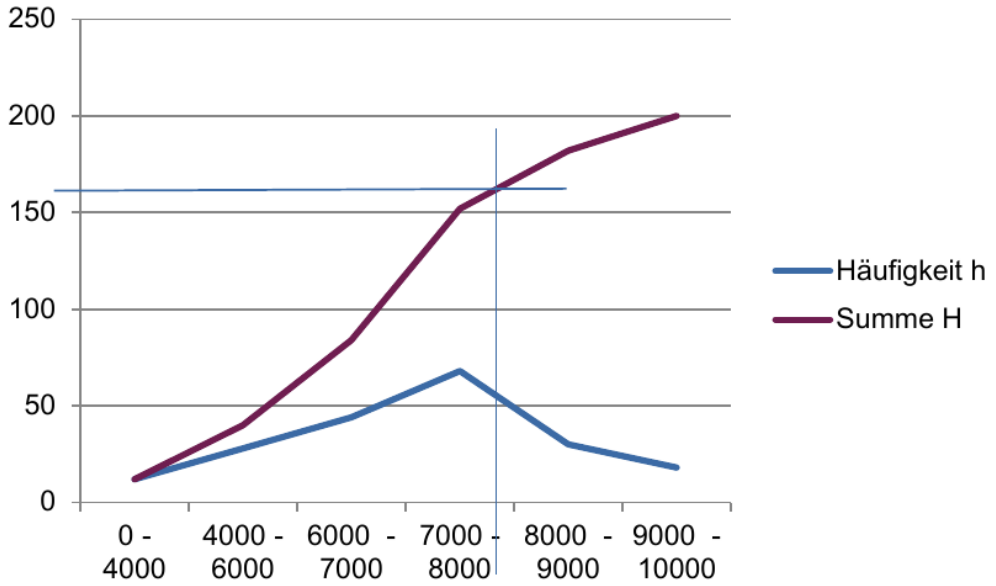
(b) $f_4 = 0.34$: 34% aller Glühbirnen haben eine Brenndauer zwischen 7000 und 8000 Stunden.

(c) $H_3 = 84$: 84 Glühbirnen haben eine Brenndauer von weniger als 7000 Stunden.

(d) $F_5 = 0.91$: 91% aller Glühbirnen haben eine Brenndauer von weniger als 9000 Stunden.

2. Erstellen Sie das Histogramm und den Polygonzug!





3. Berechnen Sie näherungsweise den Anteil der Glühbirnen mit einer Brenndauer von weniger als 6.700 Stunden.

   Ansatz: Wir verwenden eine lineare Interpolation $F(x) = b + ax$ zwischen den Kategoriengrenzen.

$$F(x < 6700) = F_2 + \frac{F_3 - F_2}{x_3^o - x_3^u}(x - x_3^u)$$

$$F(x < 6700) = 0.2 + \frac{0.42 - 0.20}{7000 - 6000}(6700 - 6000) = 0.35$$

   d) Ermitteln Sie mit Hilfe des Summenpolygons den Anteil der Glühbirnen mit einer Brenndauer von mindestens 7.800 Stunden! Überprüfen Sie Ihr Ergebnis rechnerisch.

   Ansatz: Wir berechnen erst den Anteil Glühbirnen, die weniger als 7800 Stunden brennen (siehe c) und ziehen diesen Wert dann von ab.

$$F(x < 7800) = 0.42 + \frac{.76 - .42}{8000 - 7000}(7800 - 7000) = 0.69$$

$$F(x > 7800) = 1 - F(x < 7800) = 0.31$$

4. Welche Annahme haben Sie bei Ihrer Vorgehensweise unter c) und d) unterstellt?

   Dass die Häufigkeitsdichte in den betroffenen Klassen gleich ist.

5. Wie wäre das Histogramm abzuändern, wenn bei gleichbleibenden Häufigkeiten die Obergrenze der fünften Klasse 10.000 Stunden und die Grenzen der sechsten Klasse 10.000 und 12.000 Stunden betragen hätten? Erklären Sie in diesem Zusammenhang den Begriff Häufigkeitsdichte!

Unter der Annahme, dass die Häufigkeiten gleich bleiben ändert sich sicher die Häufigkeitsdichte und die Annahme unter e) ist nicht mehr gültig. Das heisst, die relative Häufigkeit h_i darf nicht mehr genommen werden, sondern die Dichte d_i:

$$d_i = \frac{h_i}{x_i^o - x_i^u}$$

Zusatzaufgaben

Übung 5. *Gleitender Durchschnitt*
Der Polygonzug einer Verteilung wäre eigentlich stetig, da die zugrunde liegenden Variablen auch stetig sind. Da jedoch oftmals nicht ein genügend grosse Anzahl Datenpunkte vorhanden und die Klassen deshalb nicht eng genug gewählt werden können, finden sich im Polygonzug entsprechend Knicke. Eine Möglichkeit, einen Polygonzug zu glätten ist, die Häufigkeit mit Hilfe benachbarte Klassen zu interpolieren. Für eine m-gliedrige Angleichung gilt (m ungerade):

$$\bar{h_i} = \frac{1}{m} \sum_{j=-\frac{m-1}{2}}^{\frac{m-1}{2}} h_{i+j}$$

Berechnen Sie die 3-gliedrige Angleichung der folgenden Häufigkeiten und zeichnen Sie beide Polygonzüge:

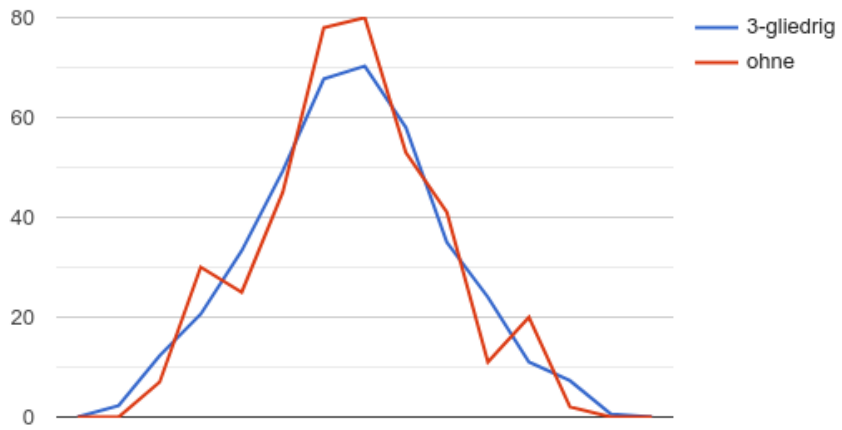
$$h_i = (0, 7, 30, 25, 45, 78, 80, 53, 41, 11, 20, 2, 0)$$

Für $m = 3$ berechnet sich der gleitende Durchschnitt $\bar{h_i}$:

$$\bar{h_i} = \frac{h_{i-1} + h_i + h_{i+1}}{3}$$

Wir benötigen zudem zwei neue Nullkategorien am Rand.

i	h_i	$\bar{h_i}$
1	0	0.0
2	0	2.3
3	7	12.3
4	30	20.6
5	25	33.3
6	45	49.3
7	78	67.7
8	80	70.3
9	53	58.0
10	41	35.0
11	11	24.0
12	20	11.0
13	2	7.3
14	0	0.6
15	0	0.0



Übung 6. *Häufigkeiten*
Sie haben eine kleine Erhebung zu den Besuchern ihrer Webseite gemacht und folgende Daten gesammelt.

Browser	System
Chrome	Android
Firefox	Windows
Firefox	Android
Safari	iOS
Safari	iOS
Chrome	Android
Edge	Windows
Chrome	Linux
Chrome	Linux
Safari	iOS
Chrome	Windows
Chrome	Linux
Edge	Windows
Firefox	Linux
Firefox	Linux

1. Berechnen Sie die absolute einfache, die absolute kumulierte, die relative einfache sowie die relative kumulierte Häufigkeit des Browsers.

Browser	h_i	H_i	f_i	F_i
Chrome	6	6	0.4	0.4
Edge	2	8	0.13	0.53
Firefox	4	12	0.27	0.8
Safari	3	15	0.2	1.0

2. Berechnen Sie die absolute einfache, die absolute kumulierte, die relative einfache sowie die relative kumulierte Häufigkeit des Systems.

System	h_i	H_i	f_i	F_i
Android	3	3	0.2	0.2
Linux	5	8	0.33	0.53
Windows	4	12	0.27	0.8
iOS	3	15	0.2	1.0