INHALT

3.5 Einzelwerte (Datenreihen) – Häufigkeitsverteilungen – Häufigkeitsklassen

Entsprechende Formeln für $\bar{x}$ und s_{n-1}: Gewichtung mit h_i Klassenmitten x^*

Histogramme mit den Dichten f_i = Rechteckhöhen: $f_i = h_i / \Delta x_i$
Funktionsgraphen für empirischen Verteilungsfunktionen F_i : stetiger Polygonzug

3.9 Wahrscheinlichkeit aus der statistischen Konvergenz: $W(A) = \lim_{n\to\infty} h_n(A)$

Wir schließen von der relative Häufigkeit h(A) auf die Wahrscheinlichkeit W(A).
Wir entwickeln analoge Formeln für Dichten:
Entsprechende Formeln für μ und σ_n: Gewichtung mit f_i

3.1 LORENZKURVE

Ein Markt für ein bestimmtes Produkt umfasst die gesamte Absatzmenge $\sum M_i$.

Der Markt werde von n Unternehmen i = 1, 2, …, n mit den Absatzmengen M_i beliefert.

Der Marktanteil des Unternehmens i ergibt sich aus $m_i = \dfrac{M_i}{\sum M_i}$

Die Unternehmen seien der Größe nach geordnet, $m_1 \le m_2 \le \ldots \le m_n$.

Wir sprechen dann von den k kleinsten Unternehmen. Statt k "kleinste" spricht man auch von den k "umsatzschwächsten", "ärmsten", "anteilsschwächsten" Merkmalsträgern.

Wir suchen ein Maß für die Konzentration K des Marktes.

Beispiel 3.1

Ein Markt werde von n =5 Unternehmen beliefert.

Jedes Unternehmen repräsentiert 1/5 = 20% der Anzahl der Unternehmen, allgemein $h_i = 1/n$.

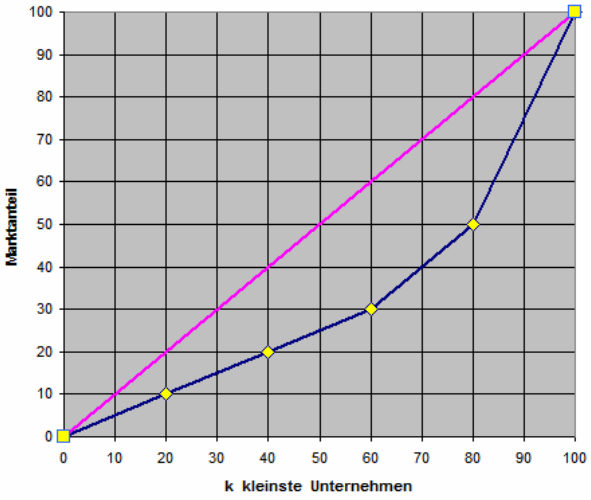
Wenn jedes Unternehmen einen Marktanteil von 20% hätte, wäre die Konzentration $K = 0$.

Aber die 3 kleinsten Unternehmen besitzen jeweils 10% Marktanteil; die restlichen beiden besitzen einen Marktanteil von 20% bzw. 50%.

Also die 3 kleinsten Unternehmen zusammen, das sind 60% der Anzahl der Unternehmen besitzen 10% + 10% + 10% = 30% der Marktanteile,

Die aufsummierten (kumulierten) Anteile lassen sich in einer Wertetabelle zusammenfassen:

Anteile einzeln h_i =1/n [%]	Markt-anteile m_i einzeln %	Anteile x_k = k/n (auf-summiert) %	Markt-anteile (auf-summiert) y_k %	Werte-paare Punkte $(x_k \mid y_k)$
		0	0	
20	10	20	10	(20 | 10)
20	10	40	20	(40 | 20)
20	10	60	30	(60 | 30)
20	20	80	50	(80 | 50)
20	50	100	100	(100 | 100)



Dem Wertepaar $(x_k \mid y_k) = (60\% \mid 30\%)$ entspricht der Punkt (60 | 30),

allgemein $(x_k \mid y_k) = \left(\dfrac{k}{n} \,\middle|\, \sum_{i=1}^{k} m_i \right) = \left(\sum_{i=1}^{k} h_i \,\middle|\, \sum_{i=1}^{k} m_i \right)$

Insgesamt ergibt sich ein Polygonzug; man nennt ihn LORENZ-Kurve.

[LORENZ, Max Otto, Wisconsin, USA, 1905]

Bei einer vollkommen gleichmäßigen Verteilung der kumulierten Häufigkeiten erhält man die Punkte (10 | 10), (20 | 20) ,…, (100 | 100). Das ergibt die "Winkelhalbierende" y = x.
Fügt man diese Gerade hinzu und ergänzt den Kurvenzug der LORENZ-Kurve durch den Punkt (0 | 0), dann erhält man ein Schaubild, das die Konzentration im Markt erkennen lässt.
Je stärker die Ungleichheit, desto stärker "hängt die LORENZ-Kurve durch".

Man spricht man von "Konzentration", wenn es um die Anteile y_k an einer Merkmalssumme in Bezug auf die Anteile x_k der Merkmalsträger geht.
Auf der x-Achse sind die kumulierten Anteile der Merkmalsträger (Personen, Unternehmen usw.) aufgetragen, auf der y-Achse die kumulierten Anteile an der gesamten Summe.

3.2 KONZENTRATIONSMAßE

Zum Messen der Konzentration K wurden Maße entwickelt, die ungefähr zwischen 0 (keine Konzentration) und 1 (totale Konzentration) liegen: $0 \le K \le 1$.
Dazu benutzt man selbstverständlich Dezimalbrüche statt Prozentzahlen. →*Excel/Konzentration*

Typ I Anteile der Merkmalsträger konstant: $h_i = \frac{1}{n}$ gleicher Anteil für alle i.

Der HERFINDAHL-Koeffizient schöpft die gesamte Information der gegebenen Verteilung aus und ist leicht berechenbar: $K_{\text{Herfindal}} = \sum_{i=1}^{n} m_i^2$ [HERFINDAHL, Orris, New York, 1950]

Er wird im Kartellrecht und in der Marktforschung verwendet.
□ $K_H = 0{,}1^2 + 0{,}1^2 + 0{,}1^2 + 0{,}2^2 + 0{,}5^2 = 0{,}32$

Der GINI-Koeffizient nutzt die Flächen unterhalb und oberhalb der LORENZ-Kurve:
Die gesamte Dreiecksfläche hat den Inhalt 0,5.
Die Fläche zwischen LORENZ-Kurve und x-Achse habe den Inhalt A_{unten}.
GINI-Koeffizient: Verhältnis der Inhalte Zwischenfläche 0,5-A : Gesamtfläche 0,5.

$$K_{\text{Gini}} = \frac{0{,}5 - A}{0{,}5} = 2(0{,}5 - A) = 1 - 2A_{\text{unten}}$$ [Gini, Corrado, Cagliari, Italien, 1910]

Die Fläche A besteht aus Trapezen gleicher Breite $h_i = 1/n$ und den Längen y_i, außerdem gilt $y_0 = 0$ und $y_n = 1$

$$A_{\text{unten}} = \left(\frac{y_0 + y_1}{2} + \frac{y_1 + y_2}{2} + \ldots + \frac{y_{n-1} + y_n}{2} \right) \cdot \frac{1}{n} = \frac{1}{n} \left(\sum_{i=0}^{n} y_i - \frac{y_0}{2} - \frac{y_n}{2} \right) = \frac{1}{n} \left(\sum_{i=1}^{n} y_i - \frac{1}{2} \right)$$

$$K_{\text{Gini}} = 1 - 2A_{\text{unten}} = 1 - \frac{2}{n} \left(\sum_{i=1}^{n} y_i - \frac{1}{2} \right)$$

→*Excel/Konzentration*

Typ II Anteile der Merkmalsträger verschieden: h_i gegeben
Der HERFINDAHL-Koeffizient findet hier keine Anwendung.
Beim GINI-Koeffizient muss man die Inhalte der Trapeze einzeln berechnen und summieren:

$$A_{\text{unten}} = \frac{y_0 + y_1}{2} \cdot h_1 + \frac{y_1 + y_2}{2} \cdot h_2 + \ldots + \frac{y_{n-1} + y_n}{2} \cdot h_n = \frac{1}{2} \sum_{i=1}^{n} (y_{i-1} + y_i) \cdot h_i$$

$K_{\text{Gini}} = 1 - 2\, A_{\text{unten}}$

Beispiel 3.2 Über die Verteilung der Vermögen in Deutschland 1995 gibt es folgende Daten:
1 % der Bevölkerung besaß 23% des Vermögens, 4.
50 % der Personen besaßen 2,5 % des Vermögens, ärmsten => P (0,5 | 0,025), (50 | 2,5)
40 % der Personen besaßen 47,5 % des Vermögens, 2.
9 % der Personen besaßen 27 % des Vermögens. 3.
Es soll die LORENZ-Kurve dargestellt und der GINI-Koeffizient berechnet werden.

1. Zunächst müssen die Anteile nach den k "ärmsten" Personen geordnet werden.
   $\frac{0{,}025}{0{,}50} = 0{,}05.$ $\frac{0{,}475}{0{,}40} = 1{,}188.$ $\frac{0{,}27}{0{,}09} = 3.$ $\frac{0{,}23}{0{,}01} = 23.$
2. Man erstellt eine Arbeitstabelle, mindestens die Wertetabelle $(x_k \mid y_k)$.
3. Mit der Wertetabelle zeichnet man das LORENZ-Polygon und die "Winkelhalbierende".
4. Dann bestimmt man den Inhalt der Fläche A und $K_{\text{Gini}} = 1 - 2\, A_{\text{unten}}$

→*Excel/Konzentration*

3.3 HERFINDAHL und GINI

Die Koeffizienten von HERFINDAHL und GINI sind die bekanntesten Konzentrationsmaße.
Sie wurden zu komplexen Instrumenten weiterentwickelt, um ihre Aussagekraft zu verbessern.
Von Strukturuntersuchungen in Gemeinden bis hin zu Weltbank und UNO werden ständig solche Koeffizienten errechnet und miteinander und im ihrer zeitlichen Entwicklung verglichen.

Jährlich werden für praktisch alle Staaten der Erde mindestens die Koeffizienten für die Einkommens- und Vermögensverteilung veröffentlicht.

Aufgabe Konzentration I – h_i konstant

gegeben: Anzahl der Merkmalsträger n, einzelne Anteile an der Merkmalssumme m_i.
oder die Mengen M_i, für zwei verschiedene Märkte

gesucht / *Schritte:*

Die Märkte sollen hinsichtlich Konzentrationsmaße verglichen werden

1. Diagramm mit "Winkelhalbierende" und beiden LORENZ-Kurven.
   nach den k anteilsschwächsten Merkmalsträgern ordnen
   Prozentangaben in Dezimalbrüche verwandeln.
   $h_i = k/n$, $m_i = M_i / \Sigma M_i$, *Wertetabellen erstellen*: x_k *und* y_k
   Wertepaare als Punkte eintragen, die LORENZ-Polygone zeichnen.
   Lineal! "Winkelhalbierende".
2. HERFINDAHL-Koeffizienten
   entsprechend der Formel
3. GINI-Koeffizienten
   entsprechend der Formel
4. Ergebnis interpretieren
   die Ergebnisse miteinander vergleichen,
   die Konzentration bei Szenario 1... ist höher, weil $G_1 > G_2$
   Bemerkung, falls ein Widerspruch zwischen $H_{Herfindahl}$ und K_{Gini} auftritt

Aufgabe Konzentration II – h_i variabel

gegeben: h_i und m_i (ein Markt)

gesucht / *Schritte:*

1. Diagramm mit "Winkelhalbierende" und die LORENZ-Kurve.
   Gegebene Anteilswerte interpretieren, Prozentzahlen in Dezimalzahlen umwandeln, nach den k anteilsschwächsten Merkmalsträgern ordnen
   Wertetabelle erstellen: x_k *und* y_k
   Wertepaare als Punkte eintragen, das LORENZ-Polygon zeichnen.
   Lineal! "Winkelhalbierende".
2. GINI-Koeffizient

$$A_{unten} = \frac{1}{2}\sum_{i=1}^{n}\left(y_{i-1} + y_i\right)\cdot h_i \quad dann \quad K_{Gini} = 1 - 2\,A_{unten}$$

Demonstration und Übung:
http://www.fernuni-hagen.de/newstatistics/applets/Lorenzkurve/Lorenzkurve.htm

3.4 HÄUFIGKEITSTABELLE

Die einfachste Art, Häufigkeiten darzustellen, ist die Strichliste.
Die Strichliste fasst die n Ergebnisse in k gleichartige Ergebnisse zusammen.

Beispiel 3.3 *nach Bleymüller S.7*

Der Inhaber eines Zeitungskiosks notiert an $n = 200$ Tagen die Anzahl x_i der an diesem Tag verkauften Exemplare der Zeitschrift "MOT". Das untersuchte Merkmal (die Zufallsvariable) X ist die Anzahl der an einem Tag verkauften "MOT".

Die Anzahl der Striche sind die absoluten Häufigkeiten n_i für die Anzahl x_i.
z.B. $n_4 = n(X=4) = 24$ [Tage]
$n(X \le 1) = 21 + 46 = 67$ [Tage]

Anzahl der verkauften Zeitungen x_i	Anzahl der Tage mit x_i verkauften Zeitungen
0	𝍸 𝍸 𝍸 𝍸 I
1	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 I
2	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 IIII
3	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍷
4	𝍸 𝍸 𝍸 𝍸 IIII
5	𝍸 𝍸
6	𝍷
7 und mehr	

Statt mit absoluten Häufigkeiten rechnet man meistens mit relativen Häufigkeiten $h_i = \frac{n_i}{n}$

Es ist: $\sum_{i=1}^{k} n_i = n$ $\sum_{i=1}^{k} h_i = 1$ $0 \le h_i \le 1$ $h_i \cdot 100\,[\%]$ prozentuale Häufigkeit

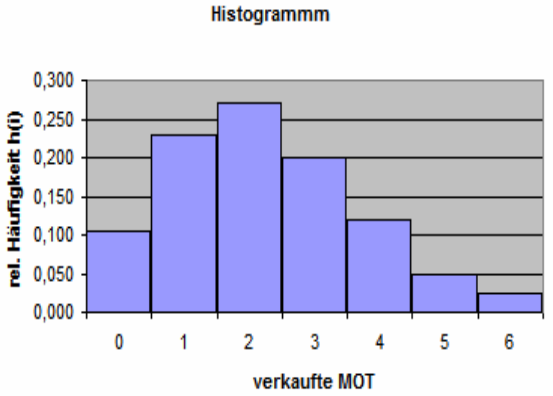
Die Funktion $h_i = h(x_i) = h(X=x_i)$ nennt man Häufigkeitsfunktion, sie ordnet jedem Ereignis i die relativen Häufigkeit h_i zu.
Der Funktionsgraph besteht aus einzelnen Punkten, man erweitert ihn zu einem Stabdiagramm (Stabhöhen ≙ h_i) oder besser zu einem Histogramm (Rechteck-Flächeninhalte = h_i, genau!).

→ *Excel / Häufigkeit*

Mit geeigneten Maßstab an der y-Achse 1 cm ≙ 0,1 (x-Achse 1cm ≙1) erhält man einen Gesamtflächen-Inhalt von 1 = 100% = 10 cm^2.

Die mittleren Ergebnisse x_i kommen am häufigsten vor, Tage mit besonders wenig Verkäufen und solche mit besonders vielen Verkäufen sind selten, das empfinden wir als normal.

→ Zentraler Grenzwertsatz, *Thema 4*
→ Normalverteilung, *Thema 5*



Wenn man die relativen Häufigkeiten aufsummiert (aufaddiert, kumuliert) erhält man die aufsummierten Häufigkeiten $F_i = h(X \le x_i)$. F_i ist die empirische Verteilungsfunktion, sie liefert die Häufigkeit dafür, dass die Zufallsvariable X höchstens das Ergebnis (die Ausprägung, den Wert, das Ereignis, den Zustand) x_i annimmt.

Funktionsgraphen, Stabdiagramme und Histogramme zu Verteilungsfunktionen F_i zeigen einen charakteristischen ∫-förmigen, monoton steigenden Verlauf mit dem Hochpunkt (x_{max} | 1), weil man normalerweise zunächst kleine h_i, dann große h_i, dann wieder kleine h_i aufsummiert.

3.5 MITTLERE WERTE

Beobachtungswerte x_i lassen sich in Häufigkeitsverteilungen übersichtlich zusammenfassen.
Mit geeigneten Maßzahlen kann man solche Häufigkeitsverteilungen noch knapper charakterisieren.
Man charakterisiert eine Datenreihe durch den Mittelwert und der Streuung um diesen Mittelwert.

1. Mittelwert $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Diese Formel gilt für Einzelwerte in Datenreihen (Thema 1 und 2).

Bei Häufigkeitsverteilungen gewichtet man mit den Häufigkeiten n_i oder h_i :

Man erhält den gewogenen Mittelwert: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i = \sum_{i=1}^{k} x_i \cdot \frac{n_i}{n} = \sum_{i=1}^{k} x_i \cdot h_i$

□ Sie kaufen 3 Sorten Äpfel ein

Mengen n_i	x_i	$x_i \cdot n_i$	h_i	h_i	$x_i \cdot h_i$
2 kg	1,80 €/kg	3,60 €	20 %	0,2	0,36
3 kg	1,20 €/kg	3,60 €	30 %	0,3	0,36
5 kg	1,00 €/kg	5,00 €	50 %	0,5	0,50
10 kg		12,20 €	100 %	1,0	1,22 €/kg

Gewogener (gewichteter) Mittelwert $\bar{x} = \frac{\sum x_i \cdot n_i}{n} = \frac{12,20}{10} = 1,22$ €/kg

2. Zentralwert (Median) x_z

□ Das Merkmal X weise n=13 Werte auf: 13; 8; 10; 15; 11; 8; 9; 11; 25 und 4 Ergebnisse liegen unter 5 Punkte. Zunächst sortiert man die Ergebnisse:

	<5	<5	<5	<5	8	8	9	10	11	11	13	15	100
F_i =	1/13	2/13	3/13	4/13	5/13	6/13	7/13	8/13 …					

→ 46,2%

→ 53,8%

Der Zentralwert $x_z = x_7 = 9$ steht in der Mitte der geordneten Beobachtungsreihe.
Allgemein: $x_z = x_i$ mit $i = 0,5 \cdot n + 0,5$

In einer Häufigkeitsverteilung ist der Zentralwert $x_z = x_i$, wobei $F_i = h(X \le x_i) = 0,5$.
Beim Ergebnis x_i wird gerade 50% der aufsummierten relativen Häufigkeit erreicht.

□ Das Merkmal Y weise n=12 Werte auf: 13; 8; 10; 15; 11; 8; 9; 11; 25 und 3 Ergebnisse liegen unter 5 Punkte. Zunächst sortiert man die Ergebnisse:

<5	<5	<5	8	8	9	||	10	11	11	13	15	25

Der Zentralwert y_z liegt zwischen $y_6 = 9$ und $y_7 = 10$ oder auch $y_z = 9,5$.
Bei großen Stichproben sind die beiden zentralen Werte $y_{n/2}$ und $y_{n/2+1}$ meistens gleich groß.

Eigenschaften des Zentralwerts (Medians)

(1) Normalerweise ergibt sich als Zentralwert ein tatsächlich existierender Wert.
(2) Der Zentralwert ist unempfindlich gegenüber "Ausreißern".
(3) Der Zentralwert lässt sich auch auf offene Randklassen (Anfang oder Ende) verwenden.
(4) Bei diskreten Zufallsvariablen ist der Zentralwert x_z sinnvoller als der Mittelwert $\bar{x}$.
Werte, die keine Metrik im physikalischen Sinne haben.
□ $x_z = 2$ Zeitschriften bzw. $\bar{x} = 2,25$ Zeitschriften

3.6 STREUUNG

Man charakterisiert eine Datenreihe durch den Mittelwert und der Streuung um diesen Mittelwert. Die Varianz ist der Mittelwert der Summe der Abweichungsquadrate.

1. Varianz für die Grundgesamtheit und für große Stichproben bei Einzelwerten:

❶ $s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \boxed{\sigma_n}^2$ ❷ $s^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \cdot N_i - \bar{x}^2$

Gewichtet mit N_i bzw. h_i : $s^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot N_i = \sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot \frac{N_i}{N} = \sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot h_i$

Dasselbe mit der Formel ❷: $s^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \cdot N_i - \bar{x}^2 = \sum_{i=1}^{k} x_i^2 \cdot h_i - \bar{x}^2$

⇒ Varianz für die Grundgesamtheit und für große Stichproben bei Häufigkeitsverteilung:

❶ $s^2 = \sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot h_i$ ❷ $s^2 = \sum_{i=1}^{k} x_i^2 \cdot h_i - \bar{x}^2$

2. Varianz für kleine Stichproben bei Einzelwerten: ($n \le 200$), n-1 Freiheitsgrade

❶ $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \boxed{\sigma_{n-1}}^2$ ❷ $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$

gewichtet mit n_i ergibt sich : $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot n_i$. $\frac{n_i}{n} = h_i$ hat hier keinen Sinn.

⇒ Varianz für kleine Stichproben bei Häufigkeitsverteilung

❶ $s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i$ ❷ $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} x_i^2 n_i - n\,\bar{x}^2\right)$

3. Die Standardabweichung $s = +\sqrt{s^2}$

4. Der Variationskoeffizient

Bei zahlenmäßig großen Merkmalen sind auch die Abweichungsquadrate größer.

- die Streuung der in einem Monat beobachteten Tagesumsätze eines Supermarktes ist viel größer als bei einem kleinen Fachgeschäft.

Die Größenordnung der Merkmale wird am ehesten durch den Mittelwert $\bar{x}$ repräsentiert.
Die Standardabweichungen verschiedener Zufallsvariablen werden vergleichbarer, wenn man sie ins Verhältnis zum Mittelwert setzt.

So kommt man zur Maßzahl Variationskoeffizient $v = \frac{s}{\bar{x}}$.

cv (coefficient of variance)

→ *Excel / Häufigkeit*

3.7 MERKMALSKLASSEN

Bei stetigen und quasi-stetigen Merkmalen x_i fasst man die x_i zu Merkmalsklassen zusammen.
Merkmale mit vielen unterschiedlichen Ausprägungen nennen wir quasi-stetige Merkmale.
Die Merkmalsklassen werden als Intervalle formuliert: $x_i \in]\, x_i^{unten} ; x_i^{oben}\,]$.
Es gibt Klassenmitten x_i^* (der Mittelwert) und Klassenbreiten $\Delta x_i = x_i^{oben} - x_i^{unten}$.

Beispiel 3.4

Bei der Untersuchung der monatlichen Bruttoverdienste von $n = 250$ Beschäftigten eines Betriebes werden $k = 10$ Klassen gebildet. Die Tabelle enthält die Bruttoverdienste x_i.

Klasse Nr.	Verdienste in €	Klassen-mitten	Beschäf-tigte Anzahl	Relative Häufigkeit en	aufsum-mierte rel. Häufigk.
i	x(i)	x_i'	n_i	$h(X = x_i)$	$h(X <= x_i)$
1	500 bis 800	650	6	0,024	0,024
2	800 < x ≤ 1100	950	13	0,052	0,076
3	1100 < x ≤ 1400	1250	22	0,088	0,164
4	1400 < x ≤ 1700	1550	32	0,128	0,292
5	1700 < x ≤ 2000	1850	40	0,160	0,452
6	2000 < x ≤ 2300	2150	42	0,168	0,620
7	] 2300 ; 2600]	2450	39	0,156	0,776
8	] 2600 ; 2900]	2750	31	0,124	0,900
9	] 2900 ; 3200]	3050	20	0,080	0,980
10	] 3200 ; 3500]	3350	5	0,020	1,000
			250	1,000	

Die Klassen sind gleich breit $\Delta x_i = 300$. Eine solche Klasseneinteilung nennt man äquidistant.
Die Rechteckflächen des Histogramms haben dann die Breite 300 und die Höhen h_i.
Man benutzt dieselben Formeln wie für Häufigkeitsverteilung ohne Klasseneinteilung.

a) Für die Merkmalsausprägungen x_i benutzt man die jeweiligen Klassenmitten x_i^*.
→ *Excel / Häufigkeit*

b) Die Binnenstrukturen innerhalb der Klassen sind unbekannt. Die relative Häufigkeiten beziehen sich auf ganze Merkmalsklassen, man spricht deshalb von Häufigkeitsdichten f_i.
Mit "Dichte" meint man, die Häufigkeit h_i sei gleichmäßig über das Intervall verteilt.
Wie in der Physik wird beim Begriff Dichte ($\rho = m/V$) die Binnenstruktur nicht betrachtet.

c) Eine Dichtefunktion f_i ordnet jedem Intervall von Merkmalen eine Häufigkeitsdichte zu.

d) Die Intervallgrenzen müssen mit den Rechteckgrenzen übereinstimmen.

e) Wenn die Klassen gleich breit sind, dann sind h_i und f_i zahlenmäßig gleich.
Man benutzt möglichst gleich-breite Klassen z.B. $\Delta x_i = 1$[cm], äquidistante Klasseneinteilung.

f) X ist eine (quasi-)stetige Zufallsvariable, die Verteilungsfunktion F_i ist stetig.
Achtung: der Funktionsgraph von F ist ein Polygonzug, keine Rechtecke!

Beispiel 3.5

Ein Lager für Gebrauchtwagen umfasst 70 PKW. Bei der Inventur ergaben sich die Werte x_i.
Bei nicht-äquidistanten Verteilungen müssen die Rechteckhöhen einzeln berechnet werden:
Für die Rechteckflächen eines Histogramms gilt:
$h_i = \Delta x_i \cdot$ Rechteckhöhe f_i.

Nr.	Wert in 1000 €	Klassen-breite in 1000 €	PKW-Anzahl	h_i
i	x(i)	Δx_i	n_i	$h(X = x_i)$
1	1 bis 2	1	8	0,114
2	2 < x ≤ 3	1	10	0,143
3	] 3 ; 4]	1	16	0,229
4	] 4 ; 5]	1	15	0,214
5	] 5 ; 7]	2	10	0,143
6	] 7 ; 9]	2	8	0,114
7	] 9 ; 15]	6	3	0,043
		14	70	1,000

Damit gilt für die Rechteckhöhen = Dichten $f_i = \dfrac{h_i}{\Delta x_i}$ → *Excel / Häufigkeit*

3.8 HÄUFIGKEITSVERTEILUNG

Aufgabe Häufigkeit

gegeben: Tabelle mit Merkmalsklassen und Häufigkeiten, Teile der Arbeitstabelle, eventuell auch Dummyspalten

Gesucht / *Schritte*:

a) Histogramm zur Häufigkeitsverteilung
   Klassenbreiten Δx_i und Klassenmitten x_i bestimmen*
   x-Achse mit den Klassengrenzen, Dichten (Rechteckhöhen) berechnen: $h_i / \Delta x_i$.
   x-Achse bezeichnen, y-Achse ohne Einheit

b) Funktionsgraph der empirischen Verteilungsfunktion.
   x-Achse mit den Klassengrenzen, $F_i = 0$ an der unteren Grenze des 1.Intervalles,
   F_i-Werte passend zu den oberen Klassengrenzen eintragen,
   Punkte mit Geradenstücken (Lineal!) verbinden. Stetig!

c) Median, Zentralwert
   das ist die Klasse oder Klassenmitte, bei der F_i 0,5 = 50% überschreitet

d) Mittelwert
   Spalte x_i h_i bilden, $\bar{x}$ ist die Spaltensumme

e) Varianz und Standardabweichung

   wenn $n \le 200$ $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} x_i^2 n_i - n\,\bar{x}^2\right)$ *wenn $n > 200$* $s^2 = \sum_{i=1}^{k} x_i^{\,2} \cdot h_i - \bar{x}^2$,

   dementsprechend benötigt man eine Spalte x_i^2 h_i oder eine Spalte x_i^2 n_i.

   Standardabweichung $\sigma = \sqrt{\sigma^2}$ *und Einheit (Benennung) hinzufügen*

f) Variationskoeffizient

   $v = \frac{s}{\bar{x}}$ [1], *keine Einheit angeben*

3.9 STATISTISCHE KONVERGENZ

a) Der Inhaber des Zeitungskiosks (3.3) fragt sich, wie groß die Wahrscheinlichkeit sein wird, dass er am nächsten Tag 3 Zeitschriften "MOT" verkaufen wird.
Der beste Schätzwert ist die gemessene relative Häufigkeit: $W(X=3) = h(X=3) = 0{,}2 = 20\%$.

b) Aus dem Lager für Gebrauchtwagen (3.5) wird ein Auto zufällig ausgewählt. Wie groß ist die Wahrscheinlichkeit, dass der Wert dieses Wagens im Bereich 5000 bis 7000 € liegt?
Der beste Schätzwert für die Wahrscheinlichkeit W ist die relative Häufigkeit:
$W(5000 < X \le 7000) = h(5000 < X \le 7000) = 0{,}143 = 14{,}3\%$.

Beispiel 3.6

In einem Zufallsversuch werden mehrmals je 40 gleichartige Reißnägel geworfen.
Es gibt zwei Elementarereignisse, "liegt auf dem Rücken" $\perp$, "Spitzenlage" $\measuredangle$.
Die Zufallsvariable X sei die Anzahl der Treffer, d.h. der Reißnägel in "Spitzenlage" $\measuredangle$.
p ist die unbekannte Treffer-Wahrscheinlichkeit.

→ *Excel / Häufigkeit*

Die relativen Häufigkeiten h_n stabilisieren sich mit zunehmendem n.
Die beste Schätzung für die Treffer-Wahrscheinlichkeit p ist für $n = 600$: $h_{600} = 0{,}393$.
$h_n = h_{600}$ liegt in einem beliebig kleinen Bereich $[p - \varepsilon\,;\, p + \varepsilon]$.
Mit zunehmendem n wird es immer wahrscheinlicher, dass h_n innerhalb dieses Bereichs bleibt.

$$\lim_{n\to\infty} W\left(\lim_{n\to\infty}(h_n - p) = 0\right) = 1$$ kurz: "h_n konvergiert statistisch zur Wahrscheinlichkeit p"

c) Mit diesen Überlegungen können wir gegenüberstellen:

empirisch beobachtete Werte	theoretisch erwartete Werte
relative Häufigkeit $h_i\ (X = x_i)$ Häufigkeitsfunktion $h(X)$	Wahrscheinlichkeit $W\ (X = x_i) = f_i = f(x_i)$ Dichtefunktion $f(X)$
aufsummierte relative Häufigkeit $h_i\ (X \le x_i)$ empirische Verteilungsfunktion $F_i = \sum_{i=1}^{k} h_i$	aufsummierte Wahrscheinlichkeit $W\ (X \le x_i)$ Verteilungsfunktion diskret: $F_i = \sum_{i=1}^{k} f_i$ stetig: $F(z) = \int_{-\infty}^{z} f(x)dx$
Mittelwert $\bar{x} = \sum_{i=1}^{k} x_i \cdot h_i$	Erwartungswert $\mu = \sum_{i=1}^{k} x_i \cdot f_i$
Varianz $s^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot h_i$ $s^2 = \sum_{i=1}^{k} x_i^2 \cdot h_i - \bar{x}^2$	erwartete Varianz $\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot f_i$ $\sigma^2 = \sum_{i=1}^{k} x_i^2 \cdot f_i - \mu^2$
Gewichtung mit relativen Häufigkeiten h_i	Gewichtung mit Wahrscheinlichkeiten f_i

d) Wenn man mit Wahrscheinlichen f_i gewichtet, spricht man von erwarteten Werten.
Für den Erwartungswert der Zufallsvariablen X schreibt man μ, μ_X oder $E(X)$ oder EX.
Für die erwartete Varianz der Zufallsvariablen X schreibt man σ^2, σ^2_X oder $V(X)$, VX.