Statistik für Wirtschaftsinformatiker
univariate Statistik - Häufigkeitsverteilung

Dr. Wolfgang Böhme

Oktober 2010

2. Univariate Verteilungsanalyse - Gliederung

- 2.0 Einführung
- **2.1 Häufigkeitsverteilung / Verteilungsfunktion**
- 2.2 Maßzahlen der Verteilung
- 2.3 Konzentrationsanalyse

2.1 Häufigkeitsverteilung / Verteilungsfunktion - Gliederung

- **2.1.1 Häufigkeitsverteilungen**
- 2.1.2 Verteilungsfunktionen

Intention

Zusammenfassung (Komprimierung) der Urliste durch Analyse der Mermalsausprägungen

Disjunkte Merkmalsausprägungen ξ_l $(l = 1, 2, ..., m)$
mit $\xi_{l_1} \neq \xi_{l_2}$ $\forall l_1, l_2$ *mit* $l_1 \neq l_2$

Merkmalstraeger	*Merkmalsauspraegung*	*disjunkte Merkmalsauspraegung*
1	x_1	ξ_1
2	x_2	ξ_1
3	x_3	ξ_2
4	x_4	ξ_1
5	x_5	ξ_2
6	x_6	ξ_3
⋮	⋮	⋮

2.1.1 Häufigkeitsverteilungen- Gliederung

- **2.1.1.1 einfache Häufigkeitsverteilungen**
- 2.1.2.2 Häufigkeitsverteilungen mit Intervallbildung

absolute Häufigkeitsverteilung

Ordnen der disjunkten Merkmalsausprägung: $\xi_{l_1} > \xi_{l_2}$ *wenn* $l_1 > l_2$
und bestimmen der absoluten Häufigkeiten n_l:

$n_l = n(\xi_l) = n(X = \xi_l)$ - Anzahl der Merkmalsausprägungen $X = \xi_l$

$$n = \sum_{l=1}^{m} n_l$$

Definition: Die absolute Häufigkeitsverteilung ist eine Zuordnung von disjunkten Merkmalsausprägungen zu deren absoluten Häufigkeiten:

$$f : \xi_l \to n_l = f(\xi_l) = n(\xi_l)$$

relative Häufigkeitsverteilung

Ausgangspunkt: absolute Häufigkeiten n_l:

$n_l = n(\xi_l)$ - Anzahl der Merkmalsausprägungen $X = \xi_l$

relative Häufigkeit p_l:

$$\begin{aligned} p_l &= p(X = \xi_l) = \frac{n_l}{n} \\ n &= \sum_{l=1}^{m} n_l \Rightarrow \\ 1 &= \sum_{l=1}^{m} p_l \end{aligned}$$

Interpretation: Anteil der Merkmalsausprägungen $(X = \xi_l)$ an der Gesamtzahl aller Merkmalsausprägungen n.

Definition: Die relative Häufigkeitsverteilung ist eine Zuordnung von disjunkten Merkmalsausprägungen zu deren relativen Häufigkeiten:

$$f : \xi_l \rightarrow p_l = f(\xi_l) = p(\xi_l)$$

Tabellendarstellung Häufigkeitsverteilung

disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*
ξ_1	n_1	p_1
ξ_2	n_2	p_2
$\vdots$	$\vdots$	$\vdots$
ξ_m	n_m	p_m

Beispiel Häufigkeitsverteilung

Datensatz: Vorlesungsbeispiel1.xls

	disjunkte Merkmalsauspraegung	*absolute Haeufigkeiten*	*relative Haeufigkeiten*
j	ξ_j	n_j	p_j
1	1	1	0.02
2	2	4	0.08
3	3	11	0.22
4	4	23	0.46
5	5	8	0.16
6	6	2	0.04
7	7	1	0.02
Summe		50	1

Grafische Darstellung der Häufigkeitsverteilung

- Balkendiagramm
- Stabdiagramm
- Kreisdiagramm

⇒ Grafische Darstellungen in Excel
(siehe Vorlesungsbeispiel1.xls)

⇒ Grafische Darstellungen in SPSS

2.1.1 Häufigkeitsverteilungen- Gliederung

- 2.1.1.1 einfache Häufigkeitsverteilungen
- **2.1.2.2 Häufigkeitsverteilungen mit Intervallbildung**

absolute Häufigkeitsverteilung mit Intervallbildung

Anstelle der disjunkten Mermalsausprägungen können disjunkte Merkmalsklassen gebildet werden: Der Wertebereich des Merkmals X einer Urliste wird durch die Werte $c_0 < c_1 < ... < c_r$ in r Intervalle unterteilt:

$$\begin{aligned} I_1 &= [c_0, c_1] \\ I_2 &= (c_1, c_2] \\ ... &\quad ... \\ I_r &= (c_{r-1}, c_r] \end{aligned}$$

Bestimmen der absoluten Häufigkeiten n_I:

$n_I = n(I_r)$ - Anzahl der Merkmalsausprägungen $X \in I_r$

Definition: Die absolute Häufigkeitsverteilung ist eine Zuordnung von disjunkten Mermalsklassen (Intervalle) zu deren absoluten Häufigkeiten:

$$f : I_r \rightarrow n_I = f(I_r) = n(I_r)$$

relative Häufigkeitsverteilung mit Intervalbildung

Ausgangspunkt: absolute Häufigkeiten n_l:

$n_l = n(I_r)$ - Anzahl der Merkmalsausprägungen $X \in I_r$

relative Häufigkeit p_l:

$$\begin{aligned} p_l &= P(X \in I_r) = \frac{n_l}{n} \\ n &= \sum_{l=1}^{m} n_l \Rightarrow \\ 1 &= \sum_{l=1}^{m} p_l \end{aligned}$$

Interpretation: Anteilder Merkmalsausprägungen ($X \in I_r$) an der gesamten Anzahlder Merkmalsausprägungen n.

Definition: Die relative Häufigkeitsverteilung ist eine Zuordnung von disjunkten Mermalsklassen (Intervalle) zu deren relativen Häufigkeiten:

$$f : I_r \rightarrow p_l = f(I_r) = p(I_r)$$

Vorlesungsbeispiel 2

Vorlesungsbeispiel 2.1-b
Statistische Masse: die einem Privathaushalt in Rechnung gestellten Telefonverbindungen
Erhebungsmerkmal: Dauer der Verbindung, in Minuten

Urliste :

1,72	2,82	3,90	4,23	1,26
0,51	7,71	1,42	1,57	0,10
0,66	1,33	026	1,06	0,39
4,14	3,51	0,76	4,60	5,42
2,92	13,13	1,99	8,17	0,77
0,14	3,23	3,86	2,59	0,27
0,97	4,28	0,87	2,15	2,14
1,57	4,32	1,20	0,38	1,19
0,10	1,33	4,32	0,04	2,02
4,41	1,29	3,11	1,62	1,74
0,09	4,68	3,64	2,58	1,05
4,67	0,97	1,07	0,51	3,15
1,68	2,13	3,60	8,05	0,23
1,93	4,41	0,12	2,01	4,59
0,27	5,39	4,11	5,98	2,28
3,11	4,52	0,55	1,78	2,38
1,79	0,48	0,16	0,97	3,47
0,87	0,22	3,13	0,34	1,70
1,54	1,15	0,46	0,63	1,74
0,11	1,03	1,86	0,44	0,84

Vorlesungsbeispiel 2

aufsteigend sortierte Urliste:

0,04	0,09	0,10	0,10	0,11
0,12	0,14	0,16	0,22	0,23
0,26	0,27	0,27	0,34	0,38
0,39	0,44	0,46	0,48	0,51
0,51	0,55	0,63	0,66	0,76
0,77	0,84	0,87	0,87	0,97
0,97	0,97	1,03	1,05	1,06
1,07	1,15	1,19	1,20	1,26
1,29	1,33	1,33	1,42	1,54
1,57	1,57	1,62	1,68	1,70
1,72	1,74	1,74	1,78	1,79
1,86	1,93	1,99	2,01	2,02
2,13	2,14	2,15	2,28	2,38
2,58	2,59	2,82	2,92	3,11
3,11	3,13	3,15	3,23	3,47
3,51	3,60	3,64	3,86	3,90
4,11	4,14	4,23	4,28	4,32
4,32	4,41	4,41	4,52	4,59
4,60	4,67	4,68	5,39	5,42
5,98	7,71	8,05	8,17	13,13

Vorlesungsbeispiel 2

Datensatz: Vorlesungsbeispiel2.xls(Telefonverbindung ; Klassenzahl m = 5)

Häufigkeitsverteilung, tabellarisch

j	x_j^u	x_j^o	n_j	p_j
1	0	2	58	0,58
2	2	4	22	0,22
3	4	6	16	0,16
4	6	10	3	0,03
m=5	10	14	1	0,01
Σ			n=100	1,00

2.1 Häufigkeitsverteilung / Verteilungsfunktion - Gliederung

- 2.1.1 Häufigkeitsverteilungen
- **2.1.2 Verteilungsfunktionen**

2.1.2 Verteilungsfunktionen- Gliederung

- **2.1.1.1 einfache Verteilungsfunktionen**
- 2.1.2.2 Verteilungsfunktionen mit Intervallbildung
- 2.1.2.3 Histogramme

absolute Summenhäufigkeitsverteilung

Ausgangspunkt: absolute Häufigkeiten n_l:

$n_l = n(\xi_l)$ - Anzahl der Merkmalsausprägungen $X = \xi_l$

absolute Summenhäufigkeit H_l:

$$H_l \quad = \quad H(X \leq \xi_l) = \sum_{r=1}^{l} n_r$$

Definition: Die **absolute Summenhäufigkeitsverteilung** ist eine Zuordnung von disjunkten Mermalsausprägungen zu den absoluten Summenhäufigkeiten:

$$f : \xi_l \rightarrow H_l = f(\xi_l) = H(X \leq \xi_l)$$

Hinweis: Merkmal X ist mindestens ordinalskaliert und die Urliste ist aufsteigend sortiert.

relative Summenhäufigkeitsverteilung

Ausgangspunkt: relative Häufigkeiten p_l:

$p_l = p(\xi_l)$ - relative Häufigkeit der Merkmalsausprägungen $X = \xi_l$
relative Summenhäufigkeit F_l:

$$F_l \quad = \quad F(X \leq \xi_l) = \frac{H(X \leq \xi_l)}{n} = \sum_{r=1}^{l} p_r$$

Definition: Die **relative Summenhäufigkeitsverteilung** ist eine Zuordnung von disjunkten Mermalsausprägungen zu den relativen Summenhäufigkeiten:

$$f : \xi_l \rightarrow F_l = f(\xi_l) = F(X \leq \xi_l)$$

Hinweis: Merkmal X ist mindestens ordinalskaliert und die Urliste ist aufsteigend sortiert.

Verteilungsfunktionen

Ausgangspunkt: relative Häufigkeiten p_l:

$p_l = p(\xi_l)$ - relative Häufigkeit der Merkmalsausprägungen $X = \xi_l$

$$F(x) = \begin{cases} 0 & \forall x : x < \xi_1 \\ F_l = \sum_{r=1}^{l} p_r & \forall x : \xi_l \leq x < \xi_{l+1} \quad l = 1, 2, ..., m-1 \\ 1 & \forall x : x \geq \xi_m \end{cases}$$

mit

x — *relle Zahl*

ξ_1 — *kleinster beobachtete Wert*

ξ_m — *groesster beobachtete Wert*

Vorlesungsbeispiel 1

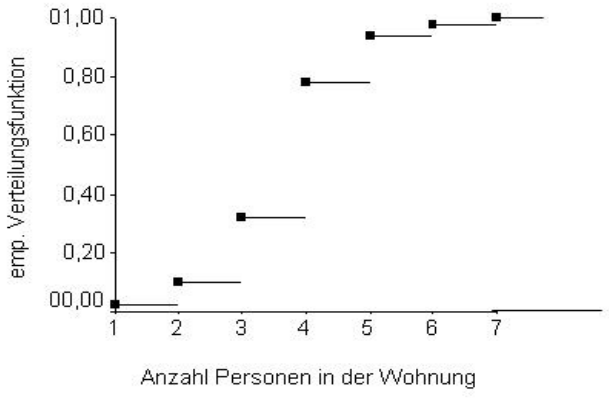
Urliste **aufsteigend sortierte Urliste**

$$\begin{pmatrix} x_1 \\ x_2 \\ . \\ x_i \\ . \\ x_n \end{pmatrix} \Rightarrow \begin{pmatrix} x_{(1)} \\ x_{(2)} \\ . \\ x_{(i)} \\ . \\ x_{(n)} \end{pmatrix} \Downarrow$$

Häufigkeitsverteilung, tabellarisch

j	ξ_j	n_j	p_j	H_j	F_j
1	1	1	0,02	1	0,02
2	2	4	0,08	5	0,10
3	3	11	0,22	16	0,32
4	4	23	0,46	39	0,78
5	5	8	0,16	47	0,94
6	6	2	0,04	49	0,98
m=7	7	1	0,02	50	1,00
Σ		n= 50	1,00		

Graph der empirischen Verteilungsfunktion



empirische Verteilungsfunktion im analytischen Ausdruck

$$F(x) = \begin{cases} 0 & \text{falls} \quad x < 1 \\ 0{,}02 & \text{falls} \quad 1 \le x < 2 \\ 0{,}10 & \text{falls} \quad 2 \le x < 3 \\ 0{,}32 & \text{falls} \quad 3 \le x < 4 \\ 0{,}78 & \text{falls} \quad 4 \le x < 5 \\ 0{,}94 & \text{falls} \quad 5 \le x < 6 \\ 0{,}98 & \text{falls} \quad 6 \le x < 7 \\ 1{,}00 & \text{falls} \quad x \ge 7 \end{cases}$$

Verteilungsfunktionen Interpretationen

- $P(a \leq X \leq b)$ - relative Häufigkeit aller Merkmalsausprägungen X, für die gilt: $a \leq X \leq b$.
- F(x) - Wert der Verteilungsfunktion an der Stelle x.

  Inhaltlich: relative Häufigkeit aller Merkmalsausprägungen X, für die gilt: $-\infty < X \leq x$. D.h. relative Summenhäufigkeit

  $P(X \leq x) = F(x)$
- $P(X < x) = P(X \leq x) - p(X = x) = F(x) - p(X = x)$
  $p(X = x)$ - relative Häufigkeit, daß X = x
- $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$
- $P(a < X \leq b) = F(b) - F(a)$
- $P(a \leq X \leq b) = F(b) - F(a) + p(X = a)$
  $p(X = x)$ - relative Häufigkeit, daß X = x
- $P(a < X < b) = F(b) - F(a) - p(X = b)$

einige Fragestellungen und Interpretationen

$$
\begin{aligned}
P(a \leq X \leq b) &= F(b) - F(a) + p(X = a) \\
P(3 \leq X \leq 5) &= F(5) - F(3) + p(X = 3) \\
P(3 \leq X \leq 5) &= 0,94 - 0,32 + 0,22 = \\
P(3 \leq X \leq 5) &= 0,84 \\
P(3 \leq X \leq 5) &= p(X = 3) + p(X = 4) + p(X = 5) = \\
P(3 \leq X \leq 5) &= 0,22 + 0,46 + 0,16 = \\
P(3 \leq X \leq 5) &= 0,84
\end{aligned}
$$

einige Fragestellungen und Interpretationen

1. F(2) = ?
2. F(2,5) = ?
3. F(4,5) = ?
4. F(4,5) - F(2,5) = ?
5. F(10) = ?
6. F(0) = ?

2.1.2 Verteilungsfunktionen- Gliederung

- 2.1.1.1 einfache Verteilungsfunktionen
- **2.1.2.2 Verteilungsfunktionen mit Intervallbildung**
- 2.1.2.3 Histogramme

absolute Summenhäufigkeitsverteilung - mit Intervalbildung

Ausgangspunkt: absolute Häufigkeiten n_l:

$n_l = n(I_l)$ - Anzahl der Merkmalsausprägungen $X \in I_l$ mit $I_l = (c_{l-1}, c_l]$

absolute Summenhäufigkeit H_l:

$$H_l \quad = \quad H(X \leq c_l) = \sum_{r=1}^{l} n_r$$

Definition: Die absolute Summenhäufigkeitsverteilung ist eine Zuordnung von Obergrenzen der disjunkten Mermalsklassen (Intervalle) zu den absoluten Summenhäufigkeiten:

$$f : c_l \rightarrow H_l = f(c_l) = H(X \leq c_l)$$

Hinweis: Merkmal X ist mindestens ordinalskaliert und die Urliste ist aufsteigend sortiert.

relative Summenhäufigkeitsverteilung - mit Intervalbildung

Ausgangspunkt: relative Häufigkeiten p_l:

$p_l = p(I_l)$ - relative Häufigkeit der Merkmalsausprägungen $X \in I_l$ mit $I_l = (c_{l-1}, c_l]$
relative Summenhäufigkeit F_l:

$$F_l \quad = \quad F(X \leq c_l) = \frac{H(X \leq c_l)}{n} = \sum_{r=1}^{l} p_r$$

Definition: Die relativ Summenhäufigkeitsverteilung ist eine Zuordnung von Obergrenzen der disjunkten Mermalsklassen (Intervalle) zu den relativen Summenhäufigkeiten:

$$f : c_l \rightarrow F_l = f(c_l) = F(X \leq c_l)$$

Hinweis: Merkmal X ist mindestens ordinalskaliert und die Urliste ist aufsteigend sortiert.

empirische Verteilungsfunktion - mit Intervalbildung

Ausgangspunkt: relative Häufigkeiten p_l:

X ist ein klassiertes kardinalskaliertes Merkmal

$p_l = p(I_r)$ - relative Häufigkeit der Merkmalsausprägungen $X \in I_l$

$$F(x) = \begin{cases} 0 & \forall x : x \leq c_0 \\ F_{l-1} + p_l^D (x - c_{l-1}) & \forall x : c_{l-1} < x \leq c_l \quad l = 1, 2, \ldots, m \\ 1 & \forall x : x > c_m \end{cases}$$

mit

$$\begin{aligned} x \quad &- \quad \textit{reelle Zahl} \\ F_{l-1} \quad &= \quad F(X \leq c_{l-1}) = \sum_{r=1}^{l-1} p_r \\ \Delta_l \quad &= \quad c_l - c_{l-1} \\ p_l^D \quad &= \quad \frac{p_l}{c_l - c_{l-1}} = \frac{p_l}{\Delta_l} \\ \Delta_l \quad &- \quad \textit{Intervallbreite} \\ p_l^D \quad &- \quad \textit{empirische Dichte (Haeufigkeitsdichte)} \end{aligned}$$

empirische Verteilungsfunktion mit Intervalbildung - Interpretation

- empirische Dichte p_l^D

  $p_l^D = \frac{p_l}{c_l - c_{l-1}} = \frac{p_l}{\Delta_l}$

- Gleichverteilung in der Klasse:

  $F_{l-1} + p_l^D(x - c_{l-1}) = F_{l-1} + \frac{p_l}{c_l - c_{l-1}}(x - c_{l-1})$

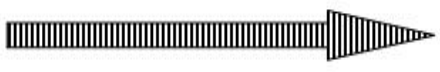
Vorlesungsbeispiel 2

Analythische Darstellung der empirischen Verteilungsfunktion (mit Intervallbildung)

Vorlesungsbeispiel 2(Telefonverbindung ; Klassenzahl m = 5)

Häufigkeitsverteilung, tabellarisch

j	x_j^u	x_j^o	n_j	p_j
1	0	2	58	0,58
2	2	4	22	0,22
3	4	6	16	0,16
4	6	10	3	0,03
m=5	10	14	1	0,01
Σ			n=100	1,00



Klassenunter- bzw. -oberarenze

Δ_j	p_j^D
2	0,2900
2	0,1100
2	0,0800
4	0,0075
4	0,0025

Vorlesungsbeispiel 2

Analythische Darstellung der empirischen Verteilungsfunktion (mit Intervallbildung)

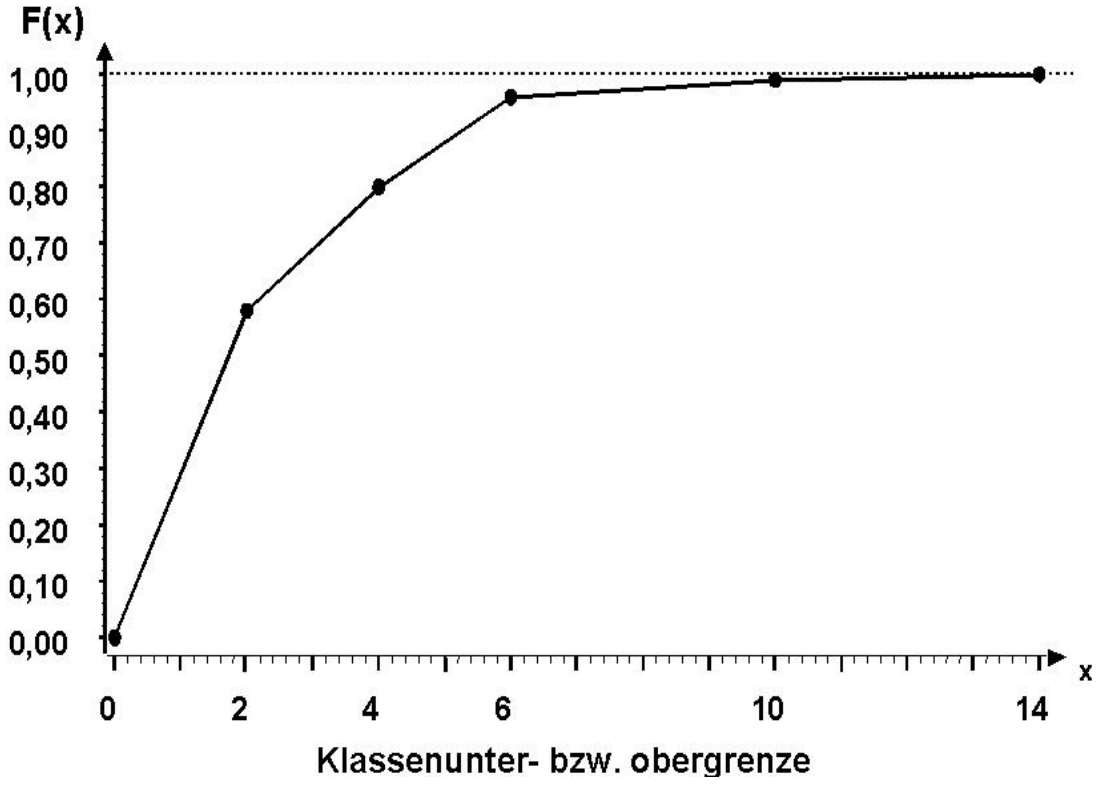
Vorlesungsbeispiel 2(Telefonverbindung ; Klassenzahl m = 5)

$$F(x)=\begin{cases} 0 & \text{falls} \quad x \le 0 \\ 0{,}29 \cdot x & \text{falls} \quad 0 < x \le 2 \\ 0{,}58 + 0{,}11 \cdot (x-2) & \text{falls} \quad 2 < x \le 4 \\ 0{,}80 + 0{,}08 \cdot (x-4) & \text{falls} \quad 4 < x \le 6 \\ 0{,}96 + 0{,}0075 \cdot (x-6) & \text{falls} \quad 6 < x \le 10 \\ 0{,}99 + 0{,}0025 \cdot (x-10) & \text{falls} \quad 10 < x \le 14 \\ 1{,}00 & \text{falls} \quad x > 14 \end{cases}$$

Vorlesungsbeispiel 2

Grafische Darstellung der empirischen Verteilungsfunktion (mit Intervallbildung)

Vorlesungsbeispiel 2(Telefonverbindung ; Klassenzahl m = 5)



Verteilungsfunktionen Interpretation

- $P(a \leq X \leq b)$ - relative Häufigkeit aller Merkmalsausprägungen X, für die gilt: $a \leq X \leq b$.
- F(x) - Wert der Verteilungsfunktion an der Stelle x.

  Inhaltlich: relative Häufigkeit aller Merkmalsausprägungen X, für die gilt: $-\infty < X \leq x$. D.h. relative Summenhäufigkeit

  $P(X \leq x) = F(x)$
- $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$
- $P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b) = F(b) - F(a)$

einige Fragestellungen und Interpretationen

$$
\begin{aligned}
P(a \leq X \leq b) &= F(b) - F(a) \\
P(3 \leq X \leq 5) &= F(5) - F(3) \\
P(3 \leq X \leq 5) &= 0,88 - 0,69 = \\
P(3 \leq X \leq 5) &= 0,19 \\
P(3 \leq X < 5) &= 0,19 \\
P(3 < X \leq 5) &= 0,19 \\
P(3 < X < 5) &= 0,19
\end{aligned}
$$

einige Fragestellungen und Interpretationen

1. F(2) = ?
2. F(2,5) = ?
3. F(4,5) = ?
4. F(4,5) - F(2,5) = ?
5. F(10) = ?
6. F(0) = ?

2.1.2 Verteilungsfunktionen- Gliederung

- 2.1.1.1 einfache Verteilungsfunktionen
- 2.1.2.2 Verteilungsfunktionen mit Intervallbildung
- **2.1.2.3 Histogramme**

Histogramme

Ein Histogramm ist die graphische Darstellung der Häufigkeitsverteilung von Messwerten. Man geht dabei von den nach Größe geordneten Daten aus und teilt den gesamten Bereich der Stichprobe in k Intervalle I_l auf. Diese müssen nicht notwendig gleich breit sein. Allerdings vereinfachen zumindest im Mittelbereich gleichgroße Klassen die Interpretation.
Über jeder Klasse wird ein Rechteck errichtet, dessen Fläche proportional zur klassenspezifischen Häufigkeit ist.
Ist die Fläche des Rechtecks gleich der absoluten Häufigkeit, wird das Histogramm absolut genannt, wenn die relativen Häufigkeiten verwendet werden, wird es entsprechend als relativ oder normiert bezeichnet.

Höhe des Rechteckes h_l:

$$\begin{aligned} \textit{absolutes} \quad &: \quad h_l = \frac{n_l}{\Delta_l} \\ \textit{relatives} \quad &: \quad h_l = \frac{n_l}{n}\frac{1}{\Delta_l} = p_l\frac{1}{\Delta_l} \\ \textit{mit } \Delta_l \quad &- \quad \textit{Intervallbreite} \end{aligned}$$

Breite des Rechteckes = Intervallbreite

Histogramme

wegen

$$
\begin{aligned}
1 &= \sum_{l=1}^{k} p_l \text{ und} \\
h_l &= \frac{n_l}{n}\frac{1}{\Delta_l} = p_l \frac{1}{\Delta_l} \Rightarrow \\
1 &= \sum_{l=1}^{k} h_l \Delta_l
\end{aligned}
$$

Für das relative Histogramm gilt: Der gesamte Flächeninhalt aller Rechtecke hat den Wert 1.

Vorlesungsbeispiel 2

Historamm der empirischen Verteilungsfunktion (mit Intervallbildung)

Vorlesungsbeispiel 2(Telefonverbindung ; Klassenzahl m = 5)

Urliste

$$\begin{pmatrix} x_1 \\ x_2 \\ . \\ x_i \\ . \\ x_n \end{pmatrix}$$

aufsteigend sortierte Urliste

$$\begin{pmatrix} x_{(1)} \\ x_{(2)} \\ . \\ x_{(i)} \\ . \\ x_{(n)} \end{pmatrix}$$

Häufigkeitsverteilung, tabellarisch

j	x_j^u	x_j^o	n_j	p_j
1	0	2	58	0,58
2	2	4	22	0,22
3	4	6	16	0,16
4	6	10	3	0,03
m=5	10	14	1	0,01
Σ			n=100	1,00

für grafische Darstellung: empirische Dichte

Δ_j	p_j^D
2	0,2900
2	0,1100
2	0,0800
4	0,0075
4	0,0025

