Willkommen zur Vorlesung Statistik (Master)

Thema dieser Vorlesung:
Verteilung diskreter Zufallsvariablen

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Thema der nächsten beiden Vorlesungen

Die nächsten beiden Vorlesungen beschäftigen sich anhand einfacher Beispiele mit folgender Frage:

Wenn aus einer bekannten Grundgesamtheit Zufallsstichproben gezogen werden – was können wir aufgrund theoretischer Überlegungen über die Ergebnisse der Stichprobenziehungen sagen?

Die Ergebnisse charakterisieren wir auf die gleiche Weise, wie wir empirisch beobachtete Daten kennzeichnen:

- Beschreibung der Verteilung durch Angabe relativer Häufigkeiten (bzw. Wahrscheinlichkeiten) für die einzelnen Werte
- Zusammenfassende Kennzeichnung durch Mittelwert und Streuung (Varianz). Achtung – der Mittelwert von Zufallsvariablen wird „Erwartungswert“ genannt (es ist der Wert, den wir aufgrund theoretischer Überlegungen im Durchschnitt erwarten).

Dies wird im folgenden anhand diskreter und stetiger Zufallsvariablen durchdekliniert.

Diskrete Zufallsvariablen

Zufallsvariablen

Unter einer Zufallsvariable versteht man eine Variable, deren Werte bzw. Ausprägungen das Ergebnis eines Zufallsvorgangs sind.

Beispiele:

- Eine Münze wird n mal geworfen – wie oft resultiert „Kopf“?
  → Analogon zur Stichprobenziehung: Aus einer Grundgesamtheit werden n Personen gezogen – wie oft resultiert „Person mit Eigenschaft A_1“ (z. B. „arm“)?
- Ein Würfel wird n mal geworfen – welche Augenzahl ergibt sich im Durchschnitt?
  → Analogon zur Stichprobenziehung: Aus einer Grundgesamtheit werden n Personen gezogen – welches Einkommen haben die Befragten im Durchschnitt?

Im ersten Fall gibt es nur eine endliche Zahl von Ausprägungen – es handelt sich um eine diskrete Zufallsvariable.

Charakterisierung einer diskreten Zufallsvariablen

Eine diskrete Zufallsvariable wird durch ihre Wahrscheinlichkeitsfunktion charakterisiert:

$$f(x) = \begin{cases} P(X = x_i) = p_i & \text{für} \quad x = x_i, i = 1,2,... \\ 0 & \text{sonst} \end{cases}$$

Die Wahrscheinlichkeitsfunktion gibt also für jede Ausprägung von X an, wie wahrscheinlich sie ist.

Beispiel: Eine Münze wird viermal geworfen

Vier Würfe:

{K, K, K, K}
{K, K, K, Z} {K, K, Z, K} {K, Z, K, K} {Z, K, K, K}
{K, K, Z, Z} {K, Z, K, Z} {Z, K, K, Z} {K, Z, Z, K} {Z, K, Z, K} {Z, Z, K, K}
{Z, Z, Z, K} {Z, Z, K, Z} {Z, K, Z, Z} {K, Z, Z, Z}
{Z, Z, Z, Z}

Wenn wir X als die Häufigkeit des Ereignisses „Zahl“ definieren, gilt folgende Wahrscheinlichkeitsfunktion:

$$f(x) = \begin{cases} 1/16 & \text{für } X = 0 \\ 4/16 = 1/4 & \text{für } X = 1 \\ 6/16 & \text{für } X = 2 \\ 4/16 & \text{für } X = 3 \\ 1/16 & \text{für } X = 4 \\ 0 & \text{sonst} \end{cases}$$

Die Verteilungsfunktion

Die Verteilungsfunktion einer Variablen gibt an, wie groß die Wahrscheinlichkeit für einen Wert $X \leq x_i$ ist.

$$F(x) = P(X \leqslant x) = \sum_{i:x_i \leqslant x} f(x_i)$$

Im Beispiel:

$$F(x) = \begin{cases} 1/16 & \text{für } X = 0 \\ 5/16 & \text{für } X = 1 \\ 11/16 & \text{für } X = 2 \\ 15/16 & \text{für } X = 3 \\ 16/16 = 1 & \text{für } X = 4 \\ 0 & \text{sonst} \end{cases}$$

Empirische Verteilung: Beispiel

Wahrscheinlichkeitsfunktion und Verteilungsfunktion können wir verstehen in Analogie zu empirischen Verteilungen.

Mathenote im Abitur

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	3	11,1	11,1	11,1
	2	7	25,9	25,9	37,0
	3	9	33,3	33,3	70,4
	4	8	29,6	29,6	100,0
	Gesamt	27	100,0	100,0	

Die Prozente entsprechen der Wahrscheinlichkeitsfunktion.

Die kumulierten Prozente entsprechen der Verteilungsfunktion: Soundsoviel Prozent haben einen Wert nicht größer als 1, soundsoviel Prozent haben einen Wert nicht größer als 2, usw.

Erwartungswert einer diskreten Zufallsvariablen

Der Erwartungswert $E(X)$ oder μ ist der theoretisch zu erwartende „Mittelwert“ der Verteilung:

$$E(X) = x_1 p_1 + \ldots + x_k p_k + \ldots = \sum_{i \geqslant 1}^{n} x_i p_i$$

Im Beispiel:

$$\begin{aligned} E(X) &= (0 \cdot 0{,}0625) + (1 \cdot 0{,}25) + (2 \cdot 0{,}375) \\ &\quad + (3 \cdot 0{,}25) + (4 \cdot 0{,}0625) \\ &= 0 + 0{,}25 + 0{,}75 + 0{,}75 + 0{,}25 \\ &= 2 \end{aligned}$$

Bei vier Münzwürfen erwarten wir also im Durchschnitt zweimal Zahl.

Varianz einer diskreten Zufallsvariablen

Die Varianz einer Zufallsvariablen ist eine Kenngröße für das Ausmaß, in dem die Werte um den Erwartungswert streuen:

$$\sigma^2 = Var(X) = (x_1 - \mu)^2 p_1 + \ldots (x_k - \mu)^2 p_k + \ldots = \sum_{i \geqslant 1} (x_i - \mu)^2 p_i$$

Im Beispiel:

$$\begin{aligned} Var(X) &= (0-2)^2 \cdot 0{,}0625 + (1-2)^2 \cdot 0{,}25 + (2-2)^2 \cdot 0{,}375 \\ &\quad + (3-2)^2 \cdot 0{,}25 + (4-2)^2 \cdot 0{,}0625 \\ &= 0{,}25 + 0{,}25 + 0 + 0{,}25 + 0{,}25 \\ &= 1 \end{aligned}$$

Binäre Merkmale: Der Bernoulli-Vorgang

Ein Bernoulli-Experiment oder Bernoulli-Vorgang ist ein einmal durchgeführter Zufallsvorgang mit Ergebnis 0 oder 1 (Nein/Ja; nicht arm/arm; ...).

Wir schreiben: $P(X = 1) = \pi$; $\quad P(X = 0) = 1 - \pi$

Wird ein Bernoulli-Vorgang (mehrfach) wiederholt, interessieren wir uns für die Häufigkeit, mit der Ereignis 1 auftritt.

Zunächst gilt: Die Wahrscheinlichkeit für eine bestimmte Stichprobenrealisierung (z.B. 1 – 0 – 0) wird berechnet als (Multiplikationstheorem!):

$$\pi^{n_1} \cdot (1 - \pi)^{n - n_1}$$

mit n = Stichprobenumfang und n_1 = Häufigkeit von 1).

Beispiele für wiederholte Bernoulli-Vorgänge

$$\pi^{n_1} \cdot (1-\pi)^{n-n_1}$$

Beispiel 1: Zwei Würfe mit Münze

0,0: $0{,}5^0 \cdot 0{,}5^2 = 0{,}25$
1,0: $0{,}5^1 \cdot 0{,}5^1 = 0{,}25$
0,1: $0{,}5^1 \cdot 0{,}5^1 = 0{,}25$
1,1: $0{,}5^2 \cdot 0{,}5^0 = 0{,}25$

Beispiel 2: Ziehen von As aus Kartenstapel (mit Zurücklegen)

0,0: $1/13^0 \cdot 12/13^2 = 1 \cdot 0{,}852 = 0{,}852$
1,0: $1/13^1 \cdot 12/13^1 = 0{,}077 \cdot 0{,}923 = 0{,}071$
0,1: $1/13^1 \cdot 12/13^1 = 0{,}077 \cdot 0{,}923 = 0{,}071$
1,1: $1/13^2 \cdot 12/13^0 = 0{,}006 \cdot 1 = 0{,}006$

Die Binomialverteilung

Die Reihenfolge der Ziehung der Elemente ist (für unsere Zwecke) irrelevant. Die Wahrscheinlichkeit eines Wertes n_1 (Häufigkeit der ‚günstigen' Ereignisse) in Stichprobe mit Umfang n kann berechnet werden nach

$$P(X = n_1) = \binom{n}{n_1} \cdot \pi^{n_1} \cdot (1-\pi)^{n-n_1}$$
$$= \frac{n!}{(n-n_1)! \cdot n_1!} \cdot \pi^{n_1} \cdot (1-\pi)^{n-n_1}$$

Bezeichnung: „Eine Variable ist $B(n, \pi)$-verteilt", kurz $X \sim B(n, \pi)$

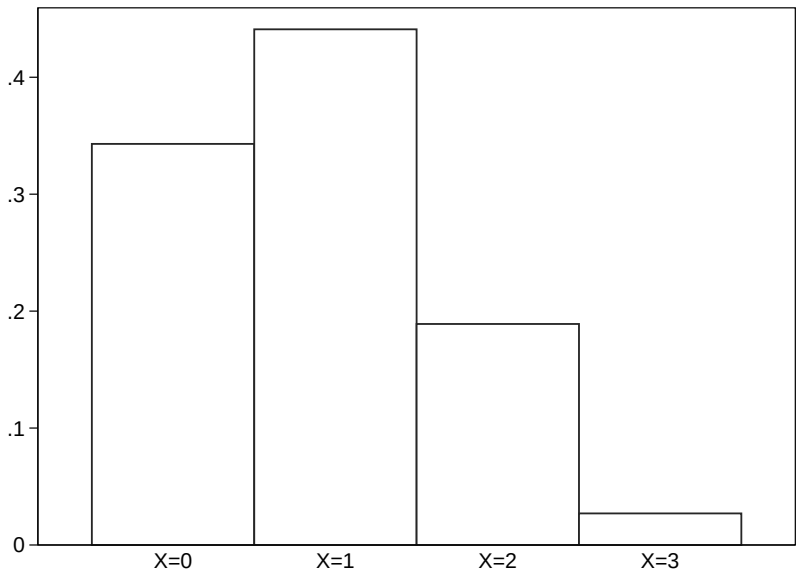
Die Binomialverteilung: Beispiel

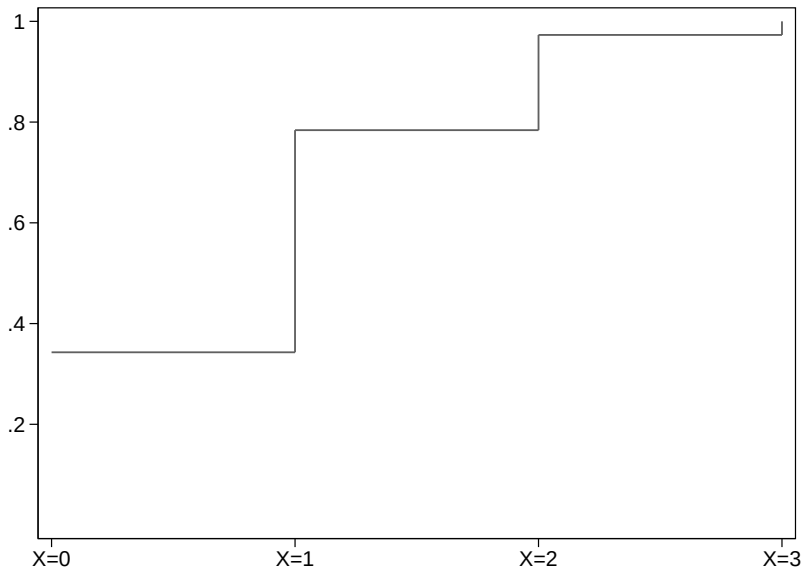
Eine Variable sei $B(3, 0{,}3)$-verteilt

$$\begin{aligned} P(X = 1) &= \binom{3}{1} \cdot (0{,}3)^1 \cdot (1 - 0{,}3)^{3-1} \\ &= \frac{3!}{2!1!} \cdot 0{,}3^1 \cdot 0{,}7^2 \\ &= \frac{3 \cdot 2 \cdot 1}{2} \cdot 0{,}3^1 \cdot 0{,}7^2 \\ &= 0{,}441 \end{aligned}$$

Weiter erhalten wir: $P(X = 0) = 0{,}343$, $P(X = 2) = 0{,}189$ und $P(X = 3) = 0{,}027$.

Die Binomialverteilung: Beispiel





Wahrscheinlichkeitsfunktion (links) und Verteilungsfunktion (rechts) lassen sich auch graphisch darstellen.

Die Binomialverteilung: E(X) und Var(X)

Im Fall der Binomialverteilung lassen sich der Erwartungswert und die Varianz besonders einfach errechnen:

$$E(X) = n \cdot \pi$$

$$Var(X) = n \cdot \pi \cdot (1 - \pi)$$

Wir erhalten in unseren Beispielen:

Vier Münzwürfe: $\mathrm{E(X)} = 4 \cdot 0{,}5 = 2$, $\mathrm{Var(X)} = 4 \cdot 0{,}5 \cdot 0{,}5 = 1$

$X \sim B(3, 0{,}3)$: $\mathrm{E(X)} = 3 \cdot 0{,}3 = 0{,}9$, $\mathrm{Var(X)} = 3 \cdot 0{,}3 \cdot 0{,}7 = 0{,}63$

Die hypergeometrische Verteilung

Für Stichproben eines binären Merkmals ohne Zurücklegen verwenden wir die hypergeometrische Verteilung $X \sim H(n, N_1, N)$:

$$f(x) = \mathrm{P}(X = n_1) = \frac{\binom{N_1}{n_1} \cdot \binom{N - N_1}{n - n_1}}{\binom{N}{n}}$$

N=Umfang der Grundgesamtheit, N_1=Anzahl der Elemente mit gesuchter Merkmalsausprägung (= Ausprägung 1), n=Gesamtumfang der Stichprobe, n_1=Teilstichprobe mit Ausprägung 1

Die hypergeometrische Verteilung: E(X) und Var(X)

Für eine Verteilung $X \sim H(n, N_1, N)$ gilt:

$$E(X) = n \cdot \frac{N_1}{N}$$

$$Var(X) = n \cdot \frac{N_1}{N} \cdot \left(1 - \frac{N_1}{N}\right) \cdot \frac{N - n}{N - 1}$$

E(X) entspricht also dem Erwartungswert einer B(n, π)-verteilten Zufallsvariablen, die Varianz ist aber um den Faktor (N-n)/(N-1) kleiner. Das liegt auch nahe, da durch das Nicht-Zurücklegen die Variabilität verringert wird. – Ist n/N klein, tendiert (N-n)/(N-1) gegen 1, kann also vernachlässigt werden.

Fazit zur diskreten Verteilungen

Wir haben uns hier ausschließlich mit einfachen Fällen befasst, die das Prinzip der Verteilung einer Zufallsvariablen verdeutlichen.

Es gibt zahlreiche andere diskrete Verteilungen, etwa für

- mehrstufige nominalskalierte Variablen, insbesondere die Multinomialverteilung, sowie
- metrische Variablen, etwa die geometrische oder die Poisson-Verteilung.

Dazu findet man in unterschiedlichen Lehrbüchern unterschiedlich detailliert Auskunft.