Klaus Pehl

Ein (Wahrscheinlichkeits-)Modell zur Relation zwischen Teilnehmenden und Teilnahmefällen in der Weiterbildung

Deutsches Institut für Erwachsenenbildung
November 2005

Online im Internet:
URL: http://www.die-bonn.de/esprid/dokumente/doc-2005/pehl05_07.pdf
Dokument aus dem Internetservice texte.online des Deutschen Instituts für Erwachsenenbildung
http://www.die-bonn.de/publikationen/online-texte/index.asp

Abstract

Klaus Pehl (2005): Ein (Wahrscheinlichkeits-)Modell zur Relation zwischen Teilnehmenden und Teilnahmefällen in der Weiterbildung

Weiterbildungseinrichtungen zählen in einem Berichtsjahr für ihr Angebot die Zahl der Teilnahmefälle. Wie groß ist der Bevölkerungsanteil in ihrem Gebiet, den sie damit erreichen? Die Beantwortung der Frage wäre ein vernünftiger Indikator im Rahmen der Weiterentwicklung der Weiterbildung nicht nur auf lokaler Ebene. Doch dafür gibt es kaum empirische Hinweise. Lediglich im Berichtssystem Weiterbildung stehen Einschätzungen von Teilnehmenden über die Zahl der von ihnen besuchten Veranstaltungen zur Verfügung. Belege von den Weiterbildungseinrichtungen fehlen bisher.
In dieser Lage kann vorläufig die Anwendung der geometrischen Verteilung, das diskrete Pendant zur Exponentialverteilung, Werte für Modellrechnungen liefern. Ihre Anwendungsmöglichkeiten für *Belegungsmodelle* – Wahrscheinlichkeitsverteilungen in einem Berichtsjahr ein, zwei oder mehr Veranstaltungen zu besuchen – werden diskutiert. Zwei Modifikationen werden für eine bessere Anpassung von Modellparametern an die vorliegenden aktuellsten Daten des Berichtssystem Weiterbildung untersucht: die Trunkierung der Verteilung auf eine maximale Zahl von Veranstaltungen und noch ergiebiger für eine Anpassung die Einführung eines Semestereffekts als Modellparameter. Nach dem in diesem Sinne besten Modell stehen hinter Belegungen zwischen 55% und 62% Individuen. Die Abschätzung des erreichten Bevölkerungsanteils aus Belegungszahlen erfordert also eine erhebliche Korrektur.

Autor

Klaus Pehl ist Leiter des Informationszentrums Weiterbildung und des Programms "Strukturwandel der Weiterbildung" am DIE.

Ein (Wahrscheinlichkeits-)Modell zur Relation zwischen Teilnehmenden und Teilnahmefällen in der Weiterbildung

Klaus Pehl, 01.09.2005

Inhalt

1 Zusammenfassung

Weiterbildungseinrichtungen zählen in einem Berichtsjahr für ihr Angebot die Zahl der Teilnahmefälle. Wie groß ist der Bevölkerungsanteil in ihrem Gebiet, den sie damit erreichen? Die Beantwortung der Frage wäre ein vernünftiger Indikator im Rahmen der Weiterentwicklung der Weiterbildung nicht nur auf lokaler Ebene. Doch dafür gibt es kaum empirische Hinweise. Lediglich im Berichtssystem Weiterbildung (BMBF 2003) stehen Einschätzungen von Teilnehmenden über die Zahl der von ihnen besuchten Veranstaltungen zur Verfügung. Belege von den Weiterbildungseinrichtungen fehlen bisher.

In dieser Lage kann vorläufig die Anwendung der geometrischen Verteilung, das diskrete Pendant zur Exponentialverteilung, Werte für Modellrechnungen liefern. Ihre Anwendungsmöglichkeiten für *Belegungsmodelle* – Wahrscheinlichkeitsverteilungen in einem Berichtsjahr ein, zwei oder mehr Veranstaltungen zu besuchen – werden diskutiert. Zwei Modifikationen werden für eine bessere Anpassung von Modellparametern an die vorliegenden aktuellsten Daten des Berichtssystem Weiterbildung untersucht: die Trunkierung der Verteilung auf eine maximale Zahl von Veranstaltungen und noch ergiebiger für eine Anpassung die Einführung eines Semestereffekts als Modellparameter. Nach dem in diesem Sinne besten Modell stehen hinter Belegungen zwischen 55% und 62% Individuen. Die Abschätzung des erreichten Bevölkerungsanteils aus Belegungszahlen erfordert also eine erhebliche Korrektur.

2 Der Kontext

Untersuchungen zur Teilnahme in der Weiterbildung werden in zwei Grundformen praktiziert. Im einen Fall werden *Individuen* aus einer Population ausgewählt und nach ihrer Weiterbildungsteilnahme in einem Bezugszeitraum befragt. Im Rahmen des für diesen Fall charakteristischen Berichtssystems Weiterbildung (Abk. BSW; BMBF 2003, 2005) wird u. a. auch die Teilnahme an Veranstaltungen der organisierten Weiterbildung differenziert nach allgemeiner und beruflicher Weiterbildung festgestellt. Im Mittelpunkt steht die Partizipation, operationalisiert als Partizipationsquote, d.h. als Anteil in % der Individuen in der Population, die an Weiterbildungsveranstaltungen im Bezugszeitraum teilgenommen haben. Auf der anderen Seite stehen die Statistiken von Weiterbildungseinrichtungen, die jährlich im Rahmen ihrer Veranstaltungsstatistik die Zahl der Teilnahme*fälle*[1] bestimmen. Charakteristisch für diese Variante ist die Statistik aus der jährlichen Vollerhebung der deutschen Volkshochschulen (Abk. VHS; Pehl/Reitz 2005). Alle berechenbaren Quoten aus der VHS-Statistik, die sich auf die Beteili-

[1] Das ist die Zahl der Belegungen; Individuen können im Berichtszeitraum eine oder *mehrere* Veranstaltungen belegt haben.

gung beziehen, sind Teilnahmefallquoten. Partizipationsquoten und Teilnahmefallquoten sind nicht mittels einfacher Berechnungen ineinander überführbar. Für zentrale Fragestellungen wäre es aber von Vorteil, auch aus Statistiken, die auf Teilnahmefällen basieren, Folgerungen für die Partizipation ziehen zu können, zumal wie im Fall der Volkshochschulen die Statistiken jährlich für einen langen Zeitraum und differenziert bis auf die kommunale Ebene vorliegen.

Die Teilnehmenden beim BSW wie die Teilnahmefälle bei der VHS-Statistik werden nach ihrem Alter in Jahren klassifiziert. Für die Betrachtung der Weiterbildung unter der Perspektive der demographischen Entwicklung ist offensichtlich, dass sie Informationen aus beiden Untersuchungsansätzen nicht isoliert, sondern aufeinander bezogen verwenden muss.

Das Grunddilemma ist, dass zwar im BSW am Rande auch Teilnahmefallzahlen erfragt werden, aber es für Weiterbildungsbetriebe wie Volkshochschulen nicht möglich ist, in der Breite ihre Fallzahlen in die Zahl von Individuen zu übersetzen. Im Rahmen der Datenbasis von Verwaltungsprogrammen werden nur exemplarische Auswertungen möglich sein. Deswegen wird hier ein anderer Weg beschritten: Sozusagen „als erster Aufschlag“ wird ein theoretisches Modell – genauer eine Familie von Modellen – präsentiert, deren Parameter mit den bereits vorliegenden Daten grob abgeschätzt werden können. Dies gibt die Möglichkeit, anhand von empirischen Daten das Modell zu überprüfen sowie es zu modifizieren und es für verschiedene Gebiete (Länder, Regionen, Kommunen) anzupassen.

3 Voraussetzungen für ein Modell

Die Gesamtzahl der Belegungen B (in einem festgelegten Gebiet in einem definierten Bezugszeitraum) setzt sich additiv aus der Zahl der Individuen zusammen, die eine Veranstaltung, zwei Veranstaltungen, drei Veranstaltungen und so fort besuchen.

Mit $p_1, p_2, p_3, ...$ seien die Wahrscheinlichkeiten $P(X = k) \quad k = 1, 2, 3, ...$ bezeichnet, dass ein Individuum 1, 2, 3, ... Veranstaltungen besucht. Dabei gelten für die $p_1, p_2, p_3, ...$ die üblichen Eigenschaften einer diskreten Wahrscheinlichkeitsverteilung

(1) $0 \le p_k \le 1$ für $k = 1, 2, 3, ...$

(2) $p_1 + p_2 + p_3 + ... = \sum_k p_k = 1 = 100\%$

Für die Gesamtzahl der Belegungen B ergibt sich eine wahrscheinlichkeitstheoretische Darstellung. Dazu werden einige Bezeichnungen eingeführt.

$G := X_1 + X_2 + X_3 + ... = \sum_k X_k$ Gesamtzahl der Individuen

$k = 1,2,3,... \quad X_k :=$ Zahl der Individuen, die k Veranstaltungen besuchen

Dann ist die Gesamtzahl der Belegungen B wegen

(3) $X_k = G \times p_k$ *für* $k = 1, 2, 3, ...$ und

(4) $B = 1 \times X_1 + 2 \times X_2 + 3 \times X_3 + ...$

darstellbar als

$$\begin{aligned} B &= 1 \times G \times p_1 + 2 \times G \times p_2 + 3 \times G \times p_3 + ... \\ &= G \times (1 \times p_1 + 2 \times p_2 + 3 \times p_3 + ...) \\ &= G \times \sum_k k \times p_k \end{aligned} \quad (5)$$

Bei gegebener Verteilung $P(X=k)=p_k \quad k=1,2,3,...$ **und Gesamtzahl der Belegungen** B **lässt sich die Gesamtzahl** G **der beteiligten Individuen ausrechnen aus**

(6) $$G=\frac{B}{\sum_k k\times p_k}=\frac{B}{E(X)}$$

$\sum_k k\times p_k$ ist der Erwartungswert $E(X)$ der Verteilung $P(X=k)=p_k \quad k=1,2,3,...$.

Mit anderen Worten: Der Zusammenhang zwischen der Gesamtzahl der Individuen und der Gesamtzahl der Belegungen ist vollständig durch die Verteilung $P(X=k)=p_k \quad k=1,2,3,...$ beschrieben. Sie soll im Weiteren als ***Belegungsmodell*** bezeichnet werden.

Folgerungen

Da die Gesamtzahl der Individuen G die Gesamtzahl der Belegungen B nicht übersteigen kann, kommen wegen (6) nur Verteilungen in Frage, deren Erwartungswert mindestens 1 beträgt.

(7) $$G\le B \Rightarrow \sum_k k\times p_k \ge 1$$

Die Gesamtzahl der Individuen G ist wegen (6) dann und nur dann gleich der Gesamtzahl der Belegungen B, wenn die Wahrscheinlichkeit *eine* Veranstaltung zu besuchen 100% und die Wahrscheinlichkeit mehr als eine Veranstaltung zu besuchen 0% ist.

(8) $G=B \Leftrightarrow p_1=1$ und $p_k=0$ *für* $k=2,3,...$

Nicht aus theoretischen Überlegungen, aber in heuristischer Betrachtung wird man sich bei der Suche nach Modellen auf solche Verteilungen beschränken, bei denen die Wahrscheinlichkeiten mit steigender Veranstaltungszahl abnehmen.

(9) $m>k \Rightarrow p_k > p_m$ oder auch

(10) $p_1 > p_2 > p_3 > ...$

Es wird sich herausstellen, dass es nicht hilft, bei den Voraussetzungen für ein geeignetes Modell bereits eine Grenze für die maximale Zahl der in einem Berichtszeitraum belegten Veranstaltungen zu nennen. Es kann sich zwar niemand vorstellen, dass größere Anzahlen als 10 wirklich vorkommen. Solche Überlegungen können aber gut nachträglich in ein Modell eingebaut werden, das zunächst ohne Begrenzung der Maximalzahl arbeitet.

4 Das Modell – ein intuitiver Zugang

Ausgangslage ist, nicht zu wissen, wie sich eine Population von Individuen bei einem Weiterbildungsangebot im Hinblick auf die Belegung von einer oder mehreren Veranstaltungen in demselben Bezugszeitraum verhalten wird. Nicht wissen heißt, keine Präferenzen zu haben und am besten vom Einfachen zum Komplexen zu gehen. Ein Modell dafür könnte so aussehen:

Tabelle 1 Der intuitive Zugang bei fehlender Information

Schritt: Teil der Population mit genau ... Veranstaltung(en)	Annahmen
1	keine Präferenz, also die Hälfte
2	keine Präferenz, also die Hälfte des Rests
3	keine Präferenz, also die Hälfte des Rests
4	keine Präferenz, also die Hälfte des Rests
usw.	usw.

Warum gerade die Hälfte? Dieser Ansatz repräsentiert in jedem Schritt eine fehlende Präferenz. Solche Überlegungen sind in der schließenden Statistik durchaus üblich. Wahrscheinlichkeitsverteilungen gehen sehr häufig unter Nutzung von Zählverfahren (Kombinatorik) von gleich wahrscheinlichen Fällen aus („Laplace-Wahrscheinlichkeiten"), was höchstens empirisch getestet werden kann, aber nicht theoretisch ableitbar ist.

Der Zugang lässt sich leicht mathematisch modellieren, wenn man sich klar macht, dass der „Rest" in jedem Schritt die nicht gewählte andere Hälfte des vorherigen Schritts darstellt.

Tabelle 2 Der intuitive Zugang mathematisch

Schritt: Teil der Population mit genau ... Veranstaltung(en)	**Berechnung des Anteils der Gesamtzahl *G* der Population**
1	50% = ½ = $\frac{1}{2^1}$
2	50% x 50% = ½ x ½ = $\frac{1}{2^2}$
3	50% x 50% x 50% = ½ x ½ x ½ = $\frac{1}{2^3}$
4 ... k	50% x 50% x 50% x 50% = ½ x ½ x ½ x ½ = $\frac{1}{2^4}$ ... $\frac{1}{2^k}$

Das intuitive Belegungsmodell wird als Wahrscheinlichkeitsverteilung beschrieben durch

(11) $P(X = k) := \frac{1}{2^k}$ *für* $k = 1, 2, 3, ...$

Dass alle Zahlen zwischen 0 und 1 liegen ist evident.

(12) $0 < \frac{1}{2^k} < 1 \, für \; k = 1,2,3,...$

Um zu erkennen, dass die Summe aller Zahlen 100% (= 1) ist, muss man in einem Analysis-Nachschlagewerk bei der Summe unendlicher Reihen nachschauen. Dort ist

(13) $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} ... = \sum_k \frac{1}{2^k} = 1$

als Beispiel angegeben.

Dort findet man auch

(14) $\frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} ... = \sum_k \frac{k}{2^k} = \sum_k k \times \frac{1}{2^k} = 2$

Für das intuitive Belegungsmodell ist also der Erwartungswert (s. 6) tatsächlich größer als 1 (s. 7). Außerdem fallen die Wahrscheinlichkeiten mit mehr belegten Veranstaltungen.

Folgerung aus dem intuitiven Belegungsmodell

Im intuitiven Belegungsmodell $P(X = r) := \frac{1}{2^r}$ *für* $r = 1, 2, 3, ...$ sind nach (6) und (14) die Belegungen (Teilnahmefallzahlen) *doppelt so hoch* wie die Gesamtzahl der Teilnehmenden.

Ist das glaubwürdig? Hätte es doch zur Folge, dass eine Volkshochschule in einer Kommune die einrichtungsspezifische Partizipationsquote als Verhältnis zwischen der teilnehmenden

Bevölkerung und der Gesamtbevölkerung [2] erheblich nach unten korrigieren müsste, wenn sie vorher aus der Zahl der Teilnahmefälle geschätzt wurde.

Letztlich kann nur die Überprüfung mit empirischen Daten Aufschluss bringen. Doch auch die theoretische Verallgemeinerung des intuitiven Belegungsmodells wird schon Möglichkeiten zur Differenzierung und für bessere Schätzungen bieten.

5 Die geometrische Verteilung als Belegungsmodell

Es wird hier nicht der Versuch unternommen, die in der einschlägigen Literatur (Upton/Cook 2004, S. 146) bekannte *geometrische Verteilung* vollständig aus dem intuitiven Belegungsmodell abzuleiten. Aber allein schon die Variation der schrittweisen Entwicklung zeigt den Weg.

Tabelle 3 Das allgemeine Belegungsmodell

Schritt: Teil der Population mit genau ... Veranstaltung(en)	**Annahmen**
1	keine Präferenz, einen Anteil p angenommen
2	keine Präferenz, also der Anteil p des Rests
3	keine Präferenz, also der Anteil p des Rests
4	keine Präferenz, also der Anteil p des Rests
usw.	usw.

Tabelle 4 Das allgemeine Belegungsmodell

Schritt: Teil der Population mit genau ... Veranstaltung(en)	**Annahme des Anteils der Gesamtzahl *G* der Population P(X = r)**
1	p
2	(1-p) x p
3	(1-(1-p) x p) x p = $(1-p)^2$ x p
4 ... k	(1-(1-(1-p) x p) x p) x p = $(1-p)^3$ x p ... $(1-p)^{k-1}$x p

Die Herleitung der allgemeinen Formel $P(X = k) = p \times (1-p)^{k-1}$ *für* $k = 1, 2, 3, ...$ bedürfte auf dem algebraischen Weg einigen Aufwand, während eine Deutung als wahrscheinlichkeitstheoretisches Experiment sie in einfacher Weise plausibel macht.

Man denke an ein Experiment mit zwei Ausgängen „Belegung“ und „Nicht-Belegung“, das unabhängig voneinander so lange wiederholt wird, bis das erste Mal „Belegung“ auftritt. Unter der Annahme, dass in jedem Versuch die Wahrscheinlichkeit für „Belegung“ p ist, sind die Wahrscheinlichkeiten, dass „Belegung“ das erste Mal im k-ten Versuch auftritt („Wartezeit k auf ersten ‚Erfolg‘“), durch die *geometrischen Verteilung*

(15) $P(X = k) = p \times (1-p)^{k-1}$ *für* $k = 1, 2, 3, ...$

gegeben.

Der Name kommt daher, dass die Wahrscheinlichkeiten eine geometrische Folge mit dem Verhältnis (1-p) bilden.

[2] Genauer definiert als der Anteil der in einem Berichtsjahr an den Veranstaltungen teilnehmenden Bevölkerung an der durchschnittlichen Bevölkerung in der Kommune im Berichtszeitraum (31.12. des Vorjahrs).

Begründung: Damit im k-ten Versuch das erste Mal „Belegung“ auftritt, muss in allen k-1 vorangegangenen Versuchen das Ereignis „Nicht-Belegung“ mit der Wahrscheinlichkeit (1-p) aufgetreten sein. Das Ergebnis ergibt sich aus der Multiplikativität von Wahrscheinlichkeiten bei Unabhängigkeit der Versuche. Das Produkt enthält k-1-mal den Faktor (1-p) und einmal den Faktor p.

Eigenschaften der geometrischen Verteilung

(16) Erwartungswert $E = \frac{1}{p}$

(17) Varianz $\sigma^2 = \frac{1-p}{p^2}$ und Standardabweichung $\sigma = \frac{\sqrt{1-p}}{p}$

(18) Der Modalwert (der Wert mit der höchsten Wahrscheinlichkeit) ist 1 für alle Werte von p.

Dass das intuitive Belegungsmodell ein Spezialfall der geometrischen Verteilung ist, ist sofort zu sehen. Man setze p = 0,5 = 50%.

(19) $P(X = k) = 0{,}5 \times (1-0{,}5)^{k-1} = 0{,}5 \times \frac{0{,}5^k}{0{,}5} = 0{,}5^k = \left(\frac{1}{2}\right)^k = \frac{1}{2^k}$

Erwartungswert und Varianz/Standardabweichung betragen

(20) Erwartungswert $E = 2$, Varianz $\sigma^2 = 2$, Standardabweichung $\sigma = \sqrt{2} \approx 1{,}41$

Die kumulativen Wahrscheinlichkeiten, die für praktische Fragestellungen häufig gebraucht werden, sind unter Benutzung der Schreibweise $q = 1 - p$:

(21) $P(X \geq k) = q^{k-1} = (1-p)^{k-1}$ und $P(X \leq k) = 1 - q^k = 1 - (1-p)^k$ (Verteilungsfunktion)

Mit der geometrischen Verteilung steht eine ganze Familie von möglichen Belegungsmodellen zur Verfügung. Der kennzeichnende Parameter ist der Anteil der Teilnehmenden, die im Berichtszeitraum genau eine Veranstaltung belegen. Er ist analog zu der Argumentation bei (6) gleichzeitig das Verhältnis zwischen Teilnehmenden und Teilnahmefällen. Der Parameter bestimmt nicht nur den Modalwert (höchste Wahrscheinlichkeit; immer bei 1; s. (18)), sondern gleichzeitig, wie schnell die Wahrscheinlichkeiten für mehr belegte Veranstaltungen absinken.

In Abbildung 1, Seite 7, und in der Tabelle 5, Seite 7, sind einige geometrische Verteilungen zum Vergleich angeboten. Zum praktischen Umgang: Wenn 70% der Teilnehmenden im Berichtszeitraum genau eine Veranstaltung belegen, dann bildet die Gesamtzahl der Teilnehmenden 70% der gezählten Teilnahmefälle.

Man liest entweder an der Abbildung 1 oder in der Tabelle 5 ab, dass der Anteil der Teilnehmenden mit zwei Veranstaltungen nur noch 21% beträgt, der mit 3 Veranstaltungen etwa bei 6% liegt und die Anteile ab 5 Veranstaltungen unter 1% bleiben.

Für das Belegungsmodell mit p = 90% ist der Anteil der Teilnehmenden die mehrere (≥ 2) Veranstaltungen belegen nur noch 10%. Für 4 oder mehr Veranstaltungen liegt der Anteil schon unter einem Promille.

Abbildung 1 Grafen ausgewählter geometrischer Verteilungen

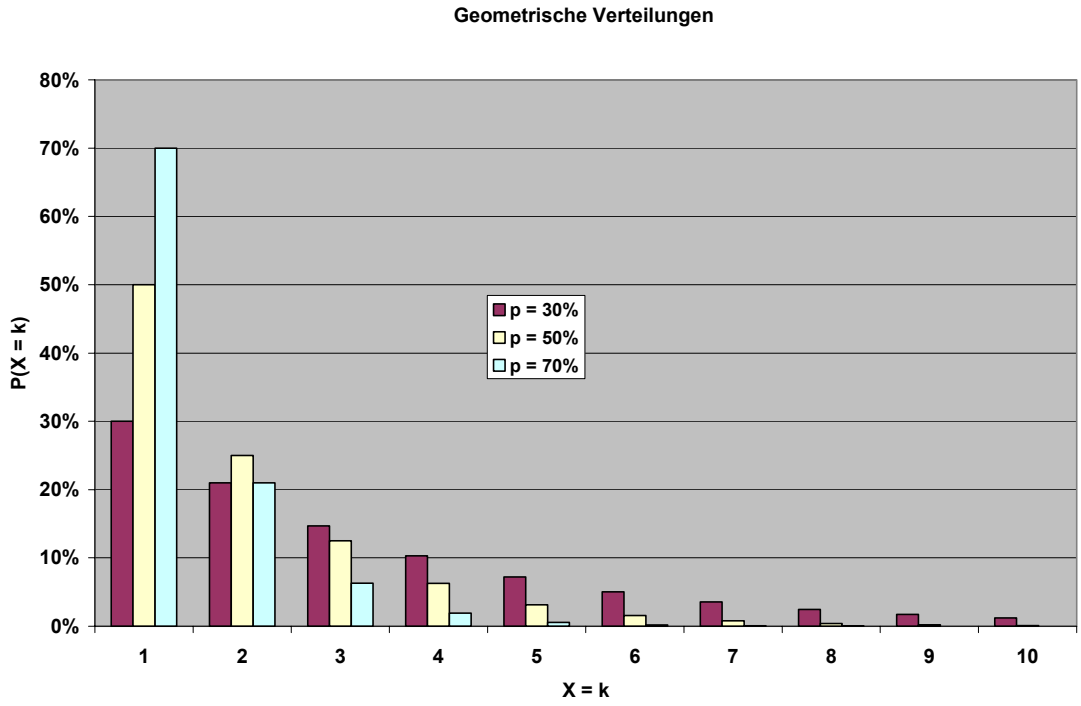


Tabelle 5 Werte und Parameter der geometrischen Verteilung P(X = k) für p = 0,1 ... 0,9 und k = 1, ...,10; k > 10

	p								
k	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
1	0,1000	0,2000	0,3000	0,4000	0,5000	0,6000	0,7000	0,8000	0,9000
2	0,0900	0,1600	0,2100	0,2400	0,2500	0,2400	0,2100	0,1600	0,0900
3	0,0810	0,1280	0,1470	0,1440	0,1250	0,0960	0,0630	0,0320	0,0090
4	0,0729	0,1024	0,1029	0,0864	0,0625	0,0384	0,0189	0,0064	0,0009
5	0,0656	0,0819	0,0720	0,0518	0,0313	0,0154	0,0057	0,0013	0,0001
6	0,0590	0,0655	0,0504	0,0311	0,0156	0,0061	0,0017	0,0003	-
7	0,0531	0,0524	0,0353	0,0187	0,0078	0,0025	0,0005	0,0001	-
8	0,0478	0,0419	0,0247	0,0112	0,0039	0,0010	0,0002	-	-
9	0,0430	0,0336	0,0173	0,0067	0,0020	0,0004	-	-	-
10	0,0387	0,0268	0,0121	0,0040	0,0010	0,0002	-	-	-
> 10	0,3487	0,1074	0,0282	0,0060	0,0010	0,0001	-	-	-
μ	**10,0**	**5,0**	**3,3**	**2,5**	**2,0**	**1,67**	**1,43**	**1,25**	**1,11**
σ	**9,49**	**4,47**	**2,79**	**1,94**	**1,41**	**1,05**	**0,78**	**0,56**	**0,35**

Die geometrische Verteilung ist die *einzige* diskrete Wahrscheinlichkeitsverteilung ***ohne Gedächtnis***, d.h.

(22) $P(X \geq n+k \mid X \geq n) = P(X = k)$ *für* $k = 1, 2, 3, \ldots$ *und* $n = 1, 2, 3, \ldots$

In Worten: Ist bekannt, dass die Zufallsgröße mindestens den Wert n hat, so ist die Wahrscheinlichkeit, dass sie diesen Wert um k übertrifft genau so groß wie die, dass sie überhaupt den Wert k annimmt. Die erwartete Wartezeit hängt nicht von der bisher verstrichenen Zeit („Vergangenheit“) ab.

Die Gedächtnislosigkeit ist eine die geometrische Verteilung definierende Eigenschaft, das heißt aus der Eigenschaft (22) lässt sich Formel (15) ableiten.

Für Mehrfachbelegungen interpretiert bedeutet die Gedächtnislosigkeit: Für jemand, der bereits n Veranstaltungen im Bezugszeitraum belegt hat, ist die Wahrscheinlichkeit eine weitere Veranstaltung zu belegen genauso groß wie die, eine erste Veranstaltung zu belegen.

6 Die trunkierte geometrische Verteilung als verbesserte Anpassung

Mit den verfügbaren Methoden ist es leichter, die geometrische Verteilung mit theoretische unendlich vielen belegbaren Veranstaltungen besser der Realität anzupassen und nur endlich viele Mehrfachbelegungen für einen Zeitraum zu betrachten. Das Verfahren heißt, die Verteilung zu *trunkieren* (Upton/Cook 2004, S. 364).

Eine Variante der tabellierten Werte (Tabelle 5, Seite 7) verdeutlicht den Vorgang. Dabei wird davon ausgegangen, dass bereits 4 die maximale Zahl von Mehrfachbelegungen ist

Tabelle 6 Werte der geometrischen Verteilung P(X = k) für p = 0,1 ... 0,9 und k = 1, ..., 4 k ≤ 4

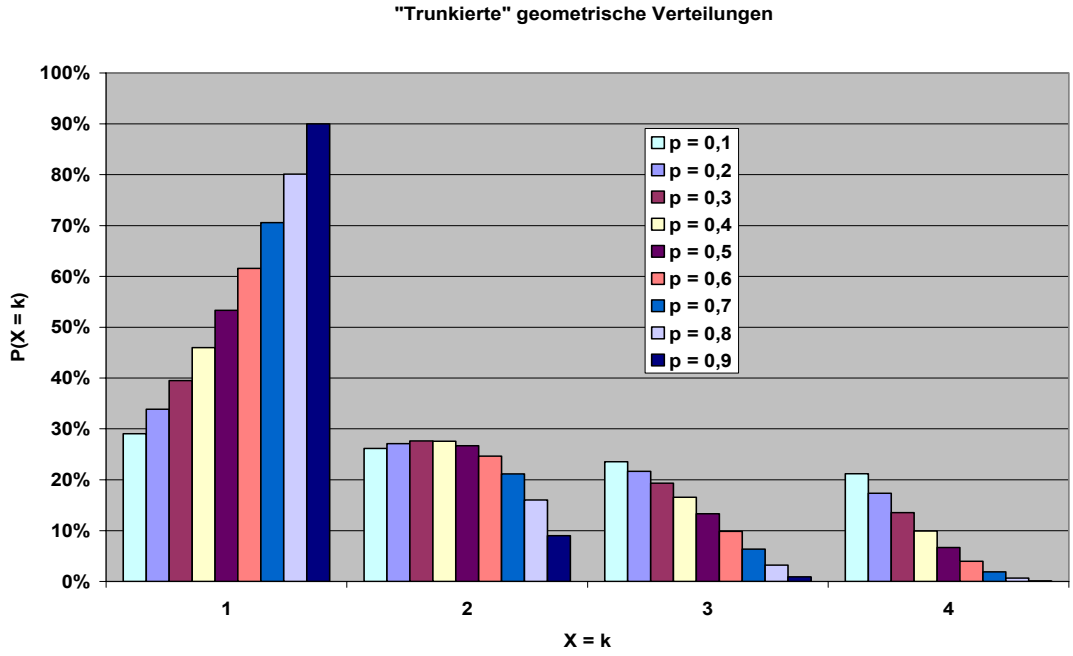
	p								
k	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
1	0,1000	0,2000	0,3000	0,4000	0,5000	0,6000	0,7000	0,8000	0,9000
2	0,0900	0,1600	0,2100	0,2400	0,2500	0,2400	0,2100	0,1600	0,0900
3	0,0810	0,1280	0,1470	0,1440	0,1250	0,0960	0,0630	0,0320	0,0090
4	0,0729	0,1024	0,1029	0,0864	0,0625	0,0384	0,0189	0,0064	0,0009
≤4	***0,3439***	***0,5904***	***0,7599***	***0,8704***	***0,9375***	***0,9744***	***0,9919***	***0,9984***	***0,9999***

Die Einzelwahrscheinlichkeiten Werte $P(X = k) \quad k = 1, \ldots, 4$ sind neu auf die Gesamtwahrscheinlichkeit $P(X \leq 4)$ zu beziehen, das heißt die bedingten Wahrscheinlichkeiten unter der Bedingung $(X \leq 4)$ auszurechnen. Es ergeben sich die trunkierten Verteilungen nach Tabelle 7 oder nach Abbildung 2.

Tabelle 7 Werte der „trunkierten“ geometrischen Verteilung P(X = k | X ≤ 4) für p = 0,1 ... 0,9

	p								
k	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
1	0,2908	0,3388	0,3948	0,4596	0,5333	0,6158	0,7057	0,8013	0,9001
2	0,2617	0,2710	0,2764	0,2757	0,2667	0,2463	0,2117	0,1603	0,0900
3	0,2355	0,2168	0,1934	0,1654	0,1333	0,0985	0,0635	0,0321	0,0090
4	0,2120	0,1734	0,1354	0,0993	0,0667	0,0394	0,0191	0,0064	0,0009

Abbildung 2 Grafen ausgewählter "trunkierter" geometrischer Verteilungen



7 Parameterschätzungen aus empirischen Werten

Für welche konkrete Verteilung aus der Familie der geometrischen Verteilungen sprechen vorliegende empirische Befunde? Wenn für eine Population in einem festgelegten Gebiet und für einen festgelegten Bezugszeitraum sowohl die Zahl der Teilnahmefälle (Belegungen) als auch die Zahl der Teilnehmenden (Individuen der Population) bekannt ist, kann wegen der Eigenschaft (6) der Parameter p geschätzt werden.

Die einzige Untersuchung, die Angaben zu beiden Größen – Teilnehmende wie Teilnahmefälle – liefert, ist das Berichtssystem Weiterbildung (BSW; BMBF 2004). In der Zusammenfassung der Ausgabe VIII für das Berichtsjahr 2000 werden empirische Verteilungen für die Zahl der belegten Veranstaltungen aufgelistet.

Tabelle 8 Verteilung und Mittelwerte der Anzahl der von Weiterbildungsteilnehmern besuchten Weiterbildungsveranstaltungen 1988-2000 aus dem Berichtssystem Weiterbildung (und Werte auf der Basis von Schätzungen für p angepasster geometrischer Verteilungen)

Zahl k der belegten Veranstaltungen	1988	1991	1994	1997	2000
1	***57% (57,0%)***	***56% (56,0%)***	***53% (53,0%)***	***55% (55,0%)***	***54% (54,0%)***
2	26% (24,5%)	28% (24,6%)	28% (24,9%)	32% (24,8%)	33% (24,8%)
3	10% (10,5%)	9% (10,8%)	10% (11,7%)	9% (11,1%)	9% (11,4%)
4	4% (4,5%)	5% (4,8%)	5% (5,5%)	3% (5,0%)	3% (5,3%)
5 oder mehr	2% (3,4%)	2% (3,8%)	4% (4,9%)	1% (4,1%)	1% (4,5%)
Mittelwerte	*1,69 (1,75)*	*1,73 (1,79)*	*1,77 (1,89)*	*1,63 (1,82)*	*1,70 (1,85)*

Quelle: BMBF 2004, S. 20 und eigene Berechnungen

Die Anpassung der geometrischen Verteilung in Tabelle 8 basiert auf der Übernahme des Werts für den Anteil für genau eine (k=1) belegte Veranstaltung als Parameter p. Hätte man eine Anpassung an der möglichst guten Übereinstimmung zwischen Erwartungswerten und Mittelwerten aus dem BSW orientiert, hätten sich die folgenden Abweichungen ergeben.

Tabelle 9 Verteilung und Mittelwerte der Anzahl der von Weiterbildungsteilnehmern besuchten Weiterbildungsveranstaltungen 1988-2000 aus dem Berichtssystem Weiterbildung (und Werte auf der Basis der Mittelwerte angepasster geometrischer Verteilungen)

Zahl der belegten Veranstaltungen	1988	1991	1994	1997	2000
1	57% (59,0%)	56% (57,8%)	53% (56,5%)	55% (61,4%)	54% (58,8%)
2	26% (24,2%)	28% (24,4%)	28% (24,6%)	32% (23,7%)	33% (24,2%)
3	10% (9,9%)	9% (10,3%)	10% (10,7%)	9% (9,2%)	9% (10,0%)
4	4% (4,0%)	5% (4,3%)	5% (4,7%)	3% (3,5%)	3% (4,1%)
5 oder mehr	2% (2,8%)	2% (3,2%)	4% (3,6%)	1% (2,2%)	1% (2,9%)
Mittelwerte	***1,69 (1,69)***	***1,73 (1,73)***	***1,77 (1,77)***	***1,63 (1,63)***	***1,70 (1,70)***

Quelle: BMBF 2004, S. 20 und eigene Berechnungen

Die Qualität der Anpassung der Modelle an die empirischen Werte erscheint zunächst zufrieden stellend wie die Tabelle der korrigierten Kontingenzkoeffizienten C* (Clauß/Finze/Partzsch 2002, S. 64) nahe legt (Tabelle 10). Die Werte, die bei starker Übereinstimmung nahe bei Null und starker Abweichung nahe bei Eins liegen müssten, überschreiten erst ab 1997 0,1.[3] Außerdem ist das Konzept, sich bei der Anpassung auf die Schätzung von p aus dem Anteil der Teilnahmefälle mit *einer* Veranstaltung zu berufen, die bessere Vorgehensweise. Einer zufallskritischen Betrachtung des „goodness of fit", wie ihn Lienert 1962 (basierend auf dem mit einem χ^2-Anpassungstest erreichten Sicherheitsniveau) vorschlägt, halten die Modelle kaum stand. Nur für 1988 und 1991 können schwache Anpassungsniveaus von 8,8% bzw. 14,0% konstatiert werden.

Tabelle 10 Korrigierte Kontingenzkoeffizienten aus der Überprüfung der Anpassung von Werten aus dem BSW und geometrischen Verteilungen (vgl. Tabelle 8 und Tabelle 9)

Kontingenzkoeffizient	1988	1991	1994	1997	2000
Basis Anpassung P(X=1)	0,0481	0,0444	0,0818	0,1444	0,1115
Basis Mittelwert = 1/p	0,0608	0,1061	0,0826	0,1906	0,2098

Der Grund für die mangelnde Güte der Anpassung kann mehrere Grunde haben. Zunächst einmal werden die angegebenen Teilnahmefallquoten im BSW ohne jede Differenzierung angegeben und beziehen sich auf die Gesamtstichprobe. Es ist bei der Inhomogenität der Bevölkerung in Weiterbildungsfragen nicht zu erwarten, dass eine einzige „Weltformel" (Modell) den Belegungsmechanismus ausreichend gut beschreiben kann. Speziell sind bei den Teilnahmefallquoten in die Veranstaltungen nicht nur solche vom Typ Kurse, Lehrgänge, Seminare, Workshops einbezogen, sondern auch Vorträge, was zu Unterschätzung von Mehrfachbelegungen durch die geometrischen Verteilungen verursachen kann. Darüber hinaus sind die Bezugszeiträume ein Jahr, was bei Weiterbildungseinrichtungen in der Regel semesterartig in zwei Arbeitsschnitte vor bzw. nach einer Sommerpause eingeteilt ist. Das führt unter Umständen zu stärkerer Repräsentanz der Belegung von *zwei* Veranstaltungen. Genau hier stellt man die stärksten Abweichungen zwischen Modell und empirischen Werten fest, besonders in den Jahren 1997 und 2000. Daher wird in einem abschließenden Abschnitt ein weiterer Schritt einer Modellanpassung erläutert, der den Gedanken des Semestereffekts als eigenen Parameter für das Modell einführt.

Zumindest liefern die geometrischen Verteilungen als Belegungsmodell einen ersten theoretischen Ansatz. Der Stichprobenumfang des BSW und die Partizipationsquoten sind so hoch, dass eine Differenzierung der Verteilung der Zahl besuchter Veranstaltungen nach allgemeiner bzw. beruflicher Weiterbildung, einem Ost-West-Vergleich, einer Differenzierung nach Altersgruppen, nach Geschlecht und anderen Merkmalen möglich wäre. Sie brächten weitere

[3] Dabei ist der Einfachheit halber für alle Jahre mit einer Zahl von Nennungen = Partizipationsquote x Stichprobenumfang des Bezugsjahrs 2000 gerechnet, als n = 0,43 * 7000.

Aufschlüsse darüber, ob und in welchen Teilbereichen sich der theoretische Ansatz bewähren könnte.

8 Geometrische Verteilungen mit „Semestereffekt“

Um die besondere Rolle zu modellieren, die eine Teilnahme an *zwei* Veranstaltungen in einem Berichtsjahr hat, werden für die entsprechende Zufallsgröße $X^{(s)}$ alle Wahrscheinlichkeiten $P(X = k)$ mit einem konstanten Faktor $1 - s$ „gestaucht“, so dass ein Zuschlag *s* („Semestereffekt“) für die Wahrscheinlichkeit $P(X^{(s)} = 2)$ möglich wird.

(23) $P(X^{(s)} = 2) = P(X = 2) \times (1 - s) + s = p \times (1 - p) \times (1 - s) + s$

Die Auswahl des Parameters ist beschränkt, denn $P(X^{(s)} = 2)$ muss eine Wahrscheinlichkeit sein und damit 1 nicht übersteigen:

(24) $0 < s < 1$

Die Randwerte $s = 1$ bzw. $s = 0$ bedeuten, dass alle Teilnehmenden genau zwei Veranstaltungen buchen bzw. dass sich kein „Semestereffekt“ auswirkt und das Modell der geometrischen Verteilung entspricht.

Mit Einführung eines Semestereffekts *s* mit $0 < s < 1$ kann die geometrische Verteilung als Belegungsmodell modifziert werden:

(25) $P(X^{(s)} = k) = p \times (1 - p)^{k-1} \times (1 - s)$ *für* $k = 1, 3, 4, ...$ **und**

(26) $P(X^{(s)} = 2) = p \times (1 - p) \times (1 - s) + s$

Die folgende Tabelle gibt einige geometrische Verteilungen mit einem Semestereffekt von *s* = 10%.

Tabelle 11 Werte der geometrischen Verteilung P(X = k) für p = 0,1 ... 0,9 mit Semestereffekt $s = 10\%$

	p								
k	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
1	0,0900	0,1800	0,2700	0,3600	0,4500	0,5400	0,6300	0,7200	0,8100
2	0,1810	0,2440	0,2890	0,3160	0,3250	0,3160	0,2890	0,2440	0,1810
3	0,0729	0,1152	0,1323	0,1296	0,1125	0,0864	0,0567	0,0288	0,0081
4	0,0656	0,0922	0,0926	0,0778	0,0563	0,0346	0,0170	0,0058	0,0008
≥ 5	0,5905	0,3686	0,2161	0,1166	0,0563	0,0230	0,0073	0,0014	0,0001

Für einen Semestereffekt von *s* = 10% ist unter den tabellierten Verteilungen die Verteilung mit dem Parameter p = 0,6 am besten an die empirischen Werte aus dem BSW, Berichtsjahr 2000 (vgl. Tabelle 8, S. 9), angepasst. Durch Variation des Parameters p findet man für verschiedene Werte des Semestereffekts die relativ beste Anpassung.

Tabelle 12 Werte des Parameters p der bestangepassten (Kontingenzkoeffizient) geometrischen Verteilung für verschiedene Semestereffekte s

	s								
	20%	**15%**	**14%**	**13%**	**12%**	**11%**	**10%**	**5%**	**1%**
p	0,6453	0,6377	0,6365	**0,6353**	0,6342	0,6331	0,6321	0,6277	0,6247
C	0,0189	0,0075	0,0068	**0,0066**	0,0069	0,0077	0,0090	0,0235	0,0444

Die beste Anpassung liegt bei der Parameterkombination p = 0,635 und einem Semestereffekt von s = 13%. Die entsprechende Verteilung zusammen mit den quadrierten gewichteten Abweichungen zu den empirischen Werten ist wie folgt tabelliert.

Tabelle 13 Bestangepasste geometrischen Verteilung mit Semestereffekt $s = 13\%$ und $p = 0{,}635$

k	P(X=k)	P(X≤k)	p_0 (BSW)	P(X=k)- p_0	$\frac{(P(X=k)-p_0)^2}{p_0}$
1	55,27%	55,27%	54,00%	0,0127	0,000299
2	33,16%	88,43%	33,00%	0,0016	0,000008
3	7,35%	95,78%	9,00%	-0,0165	0,003020
4	2,68%	98,46%	3,00%	-0,0032	0,000339
≥ 5	1,54%		1,00%	0,0054	0,002906
Summe	***100,00%***	***100,00%***	***100,00%***	***0,0000***	***0,006572***

Für diese Verteilung kann wegen der offenen Klasse „≥ 5“ ein Erwartungswert zu 1,62 nur abgeschätzt werden. Er gilt als die erwartete Anzahl der belegten Veranstaltungen in einem Berichtsjahr. Unter Verwendung von Formel (6) bedeutet dies, dass Individuen zwischen 55% und 62% einer Gruppe von Belegungen ausmachen.

Der Vollständigkeit halber sei darauf hingewiesen, dass auch die geometrischen Verteilungen mit Semestereffekt auf eine maximale Veranstaltungszahl in einem Berichtsjahr hin trunkiert werden können.

Zusätzliche Trunkierung auf einen maximale Veranstaltungszahl $k_{\max}, k_{\max} > 2,$ kombiniert mit einem Semestereffekt s mit $0 < s < 1$ modifiziert die geometrische Verteilung als Belegungsmodell zu:

(25) $P(X^{(s+t)} = k) = p \times (1-p)^{k-1} \times (1-s) / P\left(X^{(s+t)} \le k_{\max}\right)$ *für* $k = 1, 3, 4, ..., k_{\max}$ **und**

(26) $P(X^{(s+t)} = 2) = (p \times (1-p) \times (1-s) + s) / P\left(X^{(s+t)} \le k_{\max}\right)$

Für einen Semestereffekt von $s = 10\%$ sind hier einige trunkierte geometrische Verteilungen tabelliert.

Tabelle 14 Werte der trunkierten geometrischen Verteilung P(X = k) für p = 0,1 ... 0,9 und k = 1, ..., 4 k ≤ 4 mit Semestereffekt $s = 10\%$

k	p								
	0,1	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
1	0,2198	0,2851	0,3444	0,4075	0,4768	0,5527	0,6346	0,7210	0,8101
2	0,4420	0,3864	0,3687	0,3577	0,3444	0,3234	0,2911	0,2443	0,1810
3	0,1780	0,1825	0,1688	0,1467	0,1192	0,0884	0,0571	0,0288	0,0081
4	0,1602	0,1460	0,1181	0,0881	0,0597	0,0354	0,0171	0,0058	0,0008
Mittel-werte	2,28	2,19	2,06	1,92	1,76	1,61	1,46	1,32	1,20

9 Tabellenverzeichnis

10 Literatur

- Bundesministerium für Bildung und Forschung (Hrsg.) (2003): Berichtssystem Weiterbildung VIII/2000. Integrierter Gesamtbericht zur Weiterbildungssituation in Deutschland. Bonn. Online im Internet:

http://www.bmbf.de/pub/berichtssystem_weiterbildung_viii-gesamtbericht.pdf [1.9.2005]

- Bundesministerium für Bildung und Forschung (Hrsg.) (2005): Berichtssystem Weiterbildung IX. Ergebnisse der Repräsentativbefragung zur Weiterbildungssituation in Deutschland. Bonn. Online im Internet: http://www.bmbf.de/pub/berichtssystem_weiterbildung_9.pdf [1.9.2005]
- Clauß, Günter; Finze, Falk-Rüdiger; Partzsch, Lothar (2002): Statistik für Soziologen, Pädagogen, Psychologen und Mediziner. Bd.1, Grundlagen. Frankfurt am Main
- Lienert, G. A. (1962): Verteilungsfreie Methoden in der Biostatistik. Meisenheim am Glan
- Pehl, Klaus; Reitz, Gerhard (2005): Volkshochschul-Statistik. 43. Folge, Arbeitsjahr 2004. Bonn. Online im Internet: http://www.die-bonn.de/esprid/dokumente/doc-2005/pehl05_04.pdf [1.9.2005]
- Upton, Graham; Cook, Ian (2004): Oxford Dictionary of Statistics. Oxford