FEHLER BEIM ABSCHÄTZEN KUMULATIVER BINOMIAL- UND POISSON- WAHRSCHEINLICHKEITEN

nach J. GREEN & J. ROUND-TURNER, University of Liverpool
Originaltitel in "Teaching Statistics" Vol. 8 (1986), Nr. 2: The Error in Approximating Cumulative Binomial and Poisson Probabilities
Übersetzung: G. König

Zusammenfassung: Es werden Abschätzungen für die maximalen Fehler bei der Poisson-Approximation der Binomialverteilung sowie bei Approximationen der Binomial- und Poissonverteilung durch die Normalverteilung gegeben.
ZDM-Klassifikation: K60

Einleitung

Bekannt sind folgende Grenzverhalten der in der Sekundarstufe 2 im allgemeinen behandelten Verteilungsfunktionen:

- Für große Werte von n und kleine Werte von p gilt, daß eine Poissonverteilung eine gute Näherung für Binomialverteilungen ist,
- bei einer Wahrscheinlichkeit p nahe 0.5 liefert die Normalverteilung für große Werte von n gute Näherungswerte für die Binomialverteilung,
- die Normalverteilung ist eine gute Approximation für die Poissonverteilung $P(\mu)$, für große μ.

Bei solchen Approximationsaussagen und den dazugehörigen Näherungsformeln bleibt jedoch die Frage unbeantwortet, wie klein oder groß n und wie groß μ sein muß, und wie nahe p bei 0.5 liegen sollte. Was versteht man ferner unter einer guten oder ausreichenden Approximation?

Lehrbücher verweisen manchmal auf Regeln, wie z.B. $np > 5$ und $nq > 5$ $(q=1-p)$ für eine brauchbare Approximation der Binomialverteilung durch die Normalverteilung. Aber wann ist eine Approximation gut oder brauchbar? Meist fehlen auch Angaben über die Gebiete von (n,p) oder μ in denen vorgegebene Fehlerschranken der Approximationen liegen dürfen.

Im folgenden versuchen wir, die maximalen Fehler obiger Approximationen in Abhängigkeit von n und p abzuschätzen und beschreiben unsere Folgerungen durch Angabe der Approximationsformeln für diese maximalen Fehler. Unsere Folgerungen schließen die Tatsache ein, daß bei der Approximation einer Binomialverteilung durch die Poisson-Verteilung die Werte von n unbedeutend sind, der Approximationsfehler dagegen hauptsächlich von p abhängt.

Wenn wir von Verteilungen reden, dann meinen wir die kumulativen Verteilungsfunktionen (Summenwahrscheinlichkeiten), ohne im folgenden durch das zusätzliche Adjektiv kumulativ besonders darauf hinweisen zu müssen.

Die Poisson-Approximation der Binomialverteilung

Die Binomialverteilung

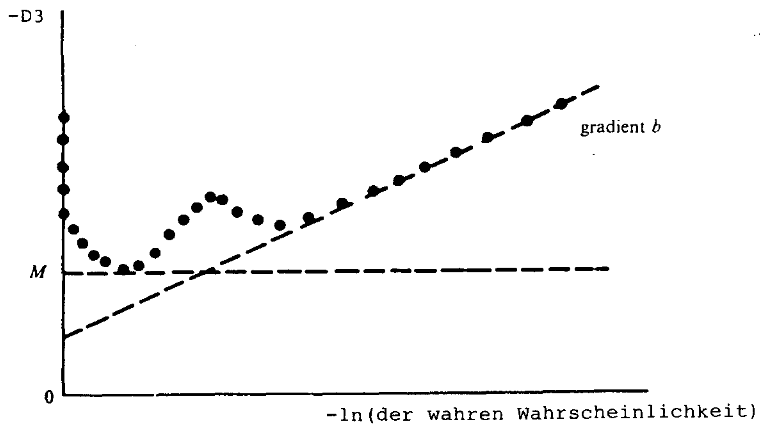
$$B(k,n,p) = \sum_{i=0}^{k} {}^{n}C_i \, p^i (1-p)^{n-i}$$

soll durch die Poisson-Verteilung

$$P(k,np) = e^{-np} \cdot \sum_{i=0}^{k} (np)^i / i!$$

approximiert werden. Wir setzen $E_1(k) = |B(k,n,p) - P(k,np)|$. $D_1(k) = \ln E_1(k)$ und $F_1(k) = \ln \{B(k,n,p)\}$. Die Funktionswerte für verschiedene k wurden bestimmt und $-D_1(k)$ gegen $-F_1(k)$ aufgetragen; dabei wurde je ein Schaubild für jeden Bereich von p und n gezeichnet. Wegen $q=1-p$ genügt es wenn wir uns auf $p < 0.5$ beschränken. Die Werte der betreffenden Funktionen wurden für $p = 0.04$ (0.04) 0.48 und n=5,10,15,20,30,40,60,80,100 berechnet; es erwies sich als nicht nötig, Schaubilder für alle Kombinationen von p und n anzufertigen. Die Schaubilder wurden für $p = 0.12$ (0.12) 0.48 und ebenso 0.04, 0.08 und obige Werte für n gezeichnet und hatten hauptsächlich die Form wie in Figur 1 dargestellt.

Wir erkennen, daß sowohl die horizontale Gerade durch den maximalen absoluten Fehler, den wir hier mit $M_1(n,p)$ bezeichnen, als auch die Tangentialgerade größer oder gleich dem Fehler für alle k Werte sind. Es ist ersichtlich, daß die Tangentengerade für die kleinen Wahrscheinlichkeiten auf der rechten und die horizontale Gerade für die größeren Wahrscheinlichkeiten auf der linken Seite schärfere Grenzen für diese Bereiche bilden. Wir werden uns hier jedoch nur mit dem absoluten maximalen Fehler M_1 beschäftigen.



Figur 1 Graph von -D3:=-ln(des absoluten Fehlers) -ln(der wahren Wahrscheinlichkeit)

p \ n	5	10	15	20	30	40	60	80	100	obere Schranke
0·04	0·003	0·005	0·007	0·007	0·007	0·007	0·006	0·006	0·006	0·008
0·08	0·011	0·015	0·015	0·013	0·013	0·012	0·011	0·012	0·011	0·016
0·12	0·021	0·023	0·018	0·019	0·018	0·017	0·017	0·017	0·017	0·024
0·24	0·048	0·041	0·039	0·037	0·036	0·036	0·036	0·035	0·035	0·048
0·36	0·058	0·062	0·060	0·059	0·058	0·058	0·056	0·056	0·056	0·072
0·48	0·095	0·072	0·085	0·085	0·083	0·082	0·082	0·081	0·081	0·096

Tabelle 1 Maximaler Fehler für die Poisson-Approximation der Binomialverteilung (exakte Werte und obere Schranke)

Die Werte von M_1 wurden für jede Kombination von n und p berechnet und in der Tabelle 1 dargestellt. Aus dieser Tabelle ist zu ersehen, daß der maximale Fehler abnimmt, wenn p abnimmt oder wenn n wächst. Wir sehen aber, daß der Einfluß von n sehr klein ist, besonders für kleine Werte von p. RAFF (1956) konnte zeigen, daß der maximale Fehler praktisch unabhängig von n ist; dies können wir bestätigen.

Die Regeln für die Anwendung der Approximation der Binomialverteilung durch die formelmäßig einfachere Poisson-Verteilung, variieren etwas in den einzelnen Lehrbüchern. In vielen Fällen wird ein zufriedenstellender Grad der Annährung für mäßig große n und mäßig kleine p angenommen, so z. B. wenn $n < 50$ und $p < 0.1$ (FRANCIS, 1979), oder $p < 0.1$ und $np < 6$ (MULHOLLAND/JONES, 1968). Weitere Vorschläge machen HOGG/TANIS (1977) mit den Bedingungen $n < 100$ und $np < 10$, bzw. CHATFIELD (1977) mit $np < 5$ und $n < 20$.

Unsere Ergebnisse zeigen, daß für $p \leq 0.12$ der maximale Fehler ≤ 0.025 für alle hier betrachteten Werte von n beträgt. Für kleinere Summenwahrscheinlichkeiten können schärfere obere Grenzen gegeben werden. So ist z. B. für $p_k < 0.001$ und $n \geq 5$ (mit $p_k = P(K \leq k)$ für die Binomialverteilung) der maximale Fehler ≤ 0.0025. Wir können sogar noch mehr sagen: Da der maximale Fehler fast unabhängig von n ist, zeichneten wir ihn in Abhängigkeit von p auf und fanden, daß die Punkte nahezu auf einer Geraden durch den Ursprung lagen. Tatsächlich ist das Verhältnis M_1/p für jeden Wert von p der Tabelle 1 kleiner oder gleich 0.2. Daraus läßt sich also folgern, daß näherungsweise $M_1 \leq 0.2p$ gilt.

Die Annäherung der Binomialverteilung durch die Normalverteilung

Nun soll B(k,n,p) durch ϕ approximiert werden, wobei ϕ die standardisierte Normalverteilung $\phi[(k+\frac{1}{2}-np)/(npq)^{1/2}]$ ist. Für jedes Paar n, p ist das Schaubild -(log des Fehlers)/ -(log der genauen Wahrscheinlichkeit) von der Form wie jenes der Figur 1, ausgenommen für einige Fälle kleinerer n und p. Wir konzentrieren uns wieder auf den maximalen Fehler, der jetzt mit $M_2(n,p)$ bezeichnet werde.

Tabelle 2 zeigt Werte der maximalen Fehler für verschiedene

Werte von n und p. Aus Symmetriegründen betrachten wir wieder keine Werte $p > 0.5$.

p \ n		5	10	15	20	30	40	60	80	100
0·04	*true*	0·062	0·100	0·094	0·076	0·051	0·053	0·041	0·033	0·030
	appr	0·111	0·090	0·078	0·070	0·058	0·050	0·039	0·032	0·026
0·08	*true*	0·094	0·071	0·047	0·050	0·039	0·031	0·027	0·023	0·020
	appr	0·093	0·075	0·065	0·058	0·049	0·042	0·033	0·026	0·021
0·16	*true*	0·060	0·042	0·033	0·026	0·023	0·020	0·016	0·014	0·012
	appr	0·063	0·050	0·043	0·039	0·032	0·028	0·022	0·018	0·014
0·24	*true*	0·031	0·026	0·022	0·018	0·014	0·013	0·011	0·010	0·008
	appr	0·034	0·026	0·023	0·020	0·017	0·014	0·011	0·009	0·007
0·32	*true*	0·026	0·015	0·014	0·012	0·010	0·0082	0·0065	0·0058	0·0051
	appr	0·022	0·016	0·014	0·012	0·010	0·0083	0·0065	0·0052	0·0042
0·40	*true*	0·013	0·0088	0·0071	0·0062	0·0050	0·0043	0·0035	0·0030	0·0027
	appr	0·011	0·0069	0·0055	0·0046	0·0036	0·0031	0·0023	0·0019	0·0015
0·48	*true*	0·0065	0·0036	0·0025	0·0019	0·0013	0·0010	0·0008	0·0006	0·0006
	appr	0·0058	0·0030	0·0020	0·0015	0·0010	0·0008	0·0006	0·0005	0·0003
0·50	*true*	0·0057	0·0027	0·0019	0·0014	0·0009	0·0007	0·0005	0·0003	0·0003
	appr	0·0056	0·0028	0·0019	0·0014	0·0009	0·0007	0·0005	0·0004	0·0003

Tabelle 2 Maximaler Fehler für die Approximation der Binomialverteilung durch die Normalverteilung (true=wahre W., appr=approximierte W.)

Die Schaubilder der Konturen von $M_2(n,p)$ und Tabelle 2 legen die Annahme nahe, daß der maximale Fehler eine wachsende Funktion von n und eine fallende Funktion von $d = |p - \frac{1}{2}|$ ist. Verschiedene funktionale Beziehungen wurden graphisch untersucht und die folgende erwies sich als die zufriedenstellendste Abschätzung

$$M_2(n,p) \simeq 0.028/n + d^2(0.7 - 0.126 \ln n)$$

Die angegebenen Werte für die Schätzungen in Tabelle 2 zeigen eine gute Übereinstimmung zwischen den wahren und den approximierten Wahrscheinlichkeiten.

Einige Beispiele für Voraussetzungen der angegebenen Approximationen aus verschiedenen Statistikbüchern sind:
Die Approximationen liefern für statistische Zwecke ausreichend genaue Werte, von $npq > 9$ (versch. Lehrbücher für die Kollegstufe) oder $0.4 < p < 0.6$ für $n > 10$ oder n sehr groß, selbst wenn p nahe bei 0 oder 1 ist. Unsere Abschätzung zeigt, daß $M_2 < 0.025$ wenn gilt

	$n > 30$	und	$0.3 < p < 0.7$,
oder	$n > 50$	und	$0.2 < p < 0.8$,
oder	$n > 100$	und	$0.1 < p < 0.9$

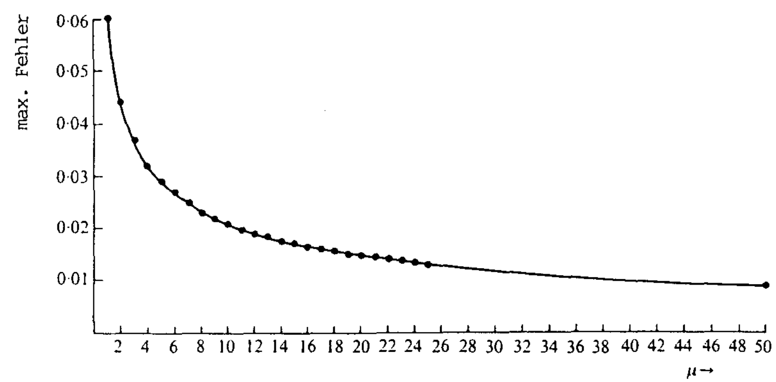
(siehe auch RAFF, 1956).

Die Approximation der Poisson-Verteilung durch die Normalverteilung

Hier haben wir

$$P_1(k,\mu) = \sum_{i=0}^{k} \frac{\mu^i e^{-\mu}}{i!} \approx \Phi\left(\frac{k + \frac{1}{2} - \mu}{\mu^{1/2}}\right)$$

Wieder konzentrieren wir uns auf den maximalen Fehler, der hier mit $M_3(\mu)$ bezeichnet werde, wobei die Maxima über die Werte von k für jedes μ bestimmt wurden. Tabelle 3 zeigt die Werte von $M_3(\mu)$ zusammen mit den angenäherten Werten. Das Schaubild von $M_3(\mu)$ ist in Figur 2 dargestellt.



Figur 2 Schaubild des maximalen Fehlers in Abhängigkeit von μ (Approximation der Poisson-Verteilung durch die Normalverteilung)

Wir stellten fest, daß das Schaubild von Logarithmus ($M_3(\mu)$) gegen Logarithmus (wahre Wahrscheinlichkeit) fast eine Gerade ergab. Damit ist die konstruierte Beziehung

$$M(\mu) \simeq 0.0627\,\mu^{-0.482}$$

sehr genau, wie aus Tabelle 3 ersehen werden kann.

μ	1	2	3	4	5	6	7	8	9
True	0·059	0·044	0·037	0·032	0·029	0·027	0·025	0·023	0·022
Approx	0·063	0·045	0·037	0·032	0·029	0·026	0·025	0·023	0·022
μ	10	11	12	13	14	15	16	17	18
True	0·0217	0·0198	0·0190	0·0183	0·0176	0·0170	0·0165	0·0160	0·0156
Approx	0·0207	0·0197	0·0189	0·0182	0·0176	0·0170	0·0165	0·0160	0·0156
μ	19	20	21	22	23	24	25	50	80
True	0·0150	0·0148	0·0144	0·0141	0·0138	0·0135	0·0132	0·0094	0·0076
Approx	0·0152	0·0148	0·0145	0·0141	0·0138	0·0136	0·0133	0·0095	0·0086

Tabelle 3 Maximaler Fehler für die Approximation der Poisson-Verteilung durch die Normalverteilung.

<u>Zusammenfassung</u>: In diesem Beitrag sind Untersuchungen über die Differenz zwischen den kumulierten Wahrscheinlichkeiten der Binomialverteilung und der sie im allgemeinen approximierenden Poisson-Verteilung beschrieben. Sieht man diese Differenzabhängigkeit von den beiden Variablen n und p, dann erweist sich, daß der Wert von n von untergeordneter Bedeutung ist.

Die Abschätzungen für die maximalen Fehler in den drei Fällen waren

$M_1(n,p) \geq 0.2p$ für die Poisson-Approximation der Binomialverteilung,

$M_2(n,p) \simeq 0.028/n + d^2(0.7 - 0.126 \ln n)$ mit $d = |p - \frac{1}{2}|$ für die Approximation der Normalverteilung an die Binomialverteilung und

$M_3(\mu) \simeq 0.0627\mu^{-0.482}$ für die Approximation der Normalverteilung an die Poisson-Verteilung.

Literatur

ALDER, H.L.; ROESSLER, E.B.: Introduction to Probability and Statistics. San Francisco: Freeman, 1977.

CHATFIELD, C.: Statistics for Technology. Middlesex: Pengiun, 1977.

FRANCIS, A.: Advanced Level Statistics. Cheltenham: Stanley Thomas, 1979.

HODGE, S.E.; SEED, M.L.: Statistics and Probability, 2. ed.. Glasgow: Blackie and Chambers, 1977.

HOGG, R.V.; TANIS, E.A.: Probability and Statistical Inference. New York: Macmillan, 1977.

LEACH, C.: Introduction to Statistics. Chichester: Wiley, 1979.

MULHOLLAND, H; JONES, C.R.: Fundamentals of Statistics. London: Butterworths, 1968.

RAFF, M.S.: On approximating the point binomial. In: J. Am. Statist. Ass., 51 (1956), S. 293 - 303.