Visualisierung stochastischer Inhalte mit GeoGebra

JÖRG MEYER, HAMELN

***Zusammenfassung**: Klassischerweise wird ein Tabellenkalkulations-Programm wie Excel verwendet, wenn man Inhalte der Schul-Stochastik visualisieren möchte.*
Wenn man mit der ästhetischen Qualität bzw. mit der Handhabbarkeit der Excel-Graphiken nicht zufrieden ist, bietet sich auch ein Geometrie-Programm wie GeoGebra an.
In diesem Beitrag wird an mehreren Beispielen demonstriert, wie man GeoGebra verwendet. Der Einsatz von GeoGebra in der Stochastik wird sich auf Demonstrationen seitens der Lehrperson beschränken.

1 Einleitung

Dass man das Geometrie-Programm GeoGebra auch gut im Stochastik-Unterricht einsetzen kann, soll an mehreren Beispielen, die inhaltlich rund um die Binomialverteilung angesiedelt sind, gezeigt werden. Dabei handelt es sich (trotz des eingebauten Zufallszahlen-Generators) weniger um Simulationen, sondern eher um die Visualisierung funktionaler Zusammenhänge und um die Auswirkungen von Parameter-Variationen.

2 Zum Histogramm der Binomialverteilung

Der Binomialverteilung liegt ein Zufallsexperiment mit genau zwei möglichen Ausgängen (Erfolg / Misserfolg) zugrunde. Die Einzelerfolgs-Wahrscheinlichkeit heißt p.

Das Zufallsexperiment wird n-mal wiederholt. X sei die Anzahl der Erfolge. Dann gilt:

$$prob(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

Das Histogramm der Binomialverteilung lässt sich – etwa auf folgende Art – gut erzeugen:

Mit Schiebereglern werden n und p definiert. Mit

prob(x) = n! / (x! *(n – x)!) * p^x * (1 – p)^(n - x)

wird die Binomialverteilung definiert (der Befehl „binomialKoeffizient“ lässt sich in GeoGebra 3.2.40.0 leider (noch) nicht funktional verwenden; das Argument muss auch (noch) „x“ heißen); man sollte bei „prob“ das Objekt nicht anzeigen lassen.

Nun braucht man nur noch das Histogramm. Die Endpunkte der Intervalle, auf denen die Balken stehen sollen, bekommt man mit

L1=Folge[k-0.5, k, 0, n+1];

die Höhen der Balken sind durch

L2=Folge[prob(k), k, 0, n]

definiert (man beachte, dass L1 ein Element mehr hat als L2).

Das Histogramm bekommt man nun vermöge

Histogramm[L1, L2];

hier wird natürlich „Objekt anzeigen“ gewählt (Abb. 1). (Es erscheint a=1; a gibt den Flächeninhalt an.)

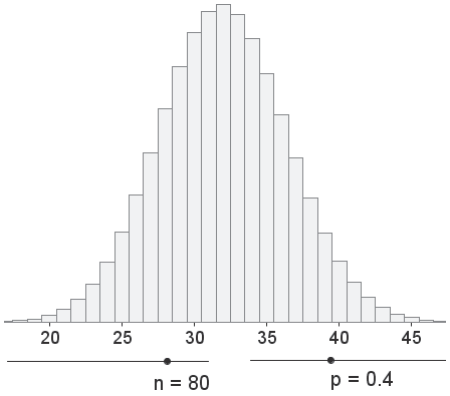

Abb. 1: Histogramm einer Binomialverteilung

Es bietet sich an, die Parameter p und n zu variieren (durch die Schieberegler) und sich qualitativ von der Gültigkeit des zentralen Grenzwertsatzes (de Moivre / Laplace) zu überzeugen.

3 Zur Normalapproximation der Binomialverteilung

Definiert man zusätzlich den Term für

$$\varphi(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-x^2/2}$$

sowie für

$$\psi(x) = \frac{1}{\sigma} \cdot \varphi\left(\frac{x-\mu}{\sigma}\right)$$

(μ ist Erwartungswert der Binomialverteilung und σ deren Standardabweichung), so ist die Normalapproximation gut zu sehen (Abb. 2). (Wieder kann man die eingebaute Funktion „normal[μ, σ, x]“ noch nicht funktional verwenden.) Man stellt

übrigens fest, dass der Graph zu *prob* fast dasselbe Aussehen wie der Graph zu ψ hat.

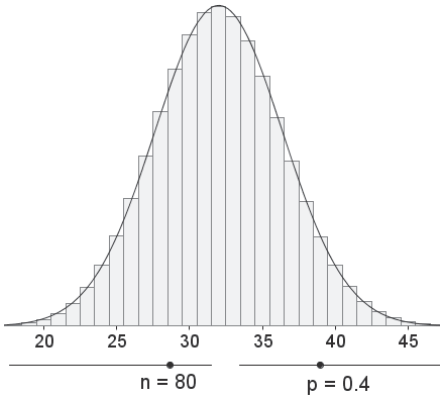

Abb. 2: Normalapproximation der Binomialverteilung

4 Zur „Umkehrung“ der Binomialverteilung

Der Erwartungswert μ ist nicht immer eine natürlich Zahl. Rundet man μ zum nächsten Ganzen, so soll das Ergebnis *m* heißen. (Ist $\mu = 2{,}5$, so ist $m = 3$.) Mitunter möchte man für eine vorgegebene Sicherheits-Wahrscheinlichkeit δ ein ganzzahliges g so bestimmen, dass

$$prob\left(m - g \le X \le m + g\right) = \delta$$

bzw.

$$\sum_{k=m-g}^{m+g} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \delta$$

gilt. Das kann man in GeoGebra gut (mit m=round(μ)) anschaulich lösen:

Für *g* führt man einen Schieberegler (mit Schrittweite 1) ein und definiert

L3=Folge[k-0.5, k, m-g, m+g+1]

und

L4=Folge[prob(k), k, m-g, m+g]

sowie

Histogramm[L3, L4].

Mit „Text einfügen“ kann man den Wert dieser neuen Teilsumme verfolgen, wenn *g* geändert wird (Abb. 3).

Man erlebt anschaulich, dass sich im Allgemeinen kein *g* so finden lässt, dass δ exakt erreicht wird.

5 Zur „Umkehrung“ der Normalverteilung

Einfacher und übersichtlicher sind die Verhältnisse bei der Normalverteilung.

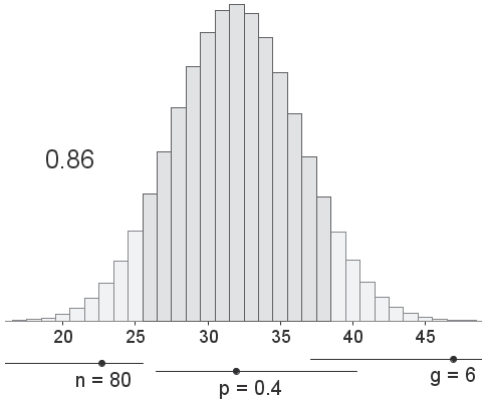

Abb. 3: Zur „Umkehrung“ der Binomialverteilung

Mit $\varphi(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-x^2/2}$ und $\int_{-z}^{z} \varphi(x) \cdot \mathrm{d}x = \delta$ hat man einen eineindeutigen Zusammenhang zwischen *z* und δ.

Dabei ist der Weg von *z* nach δ vermöge

Integral[φ, -z, z] trivial.

Der umgekehrte Weg lässt sich durch Variation des dicken Punktes A auf der Rechtsachse anschaulich machen (Abb. 4); man bekommt δ aus

Integral[φ, -x(A), x(A)]. Natürlich ist $z = x(A)$.

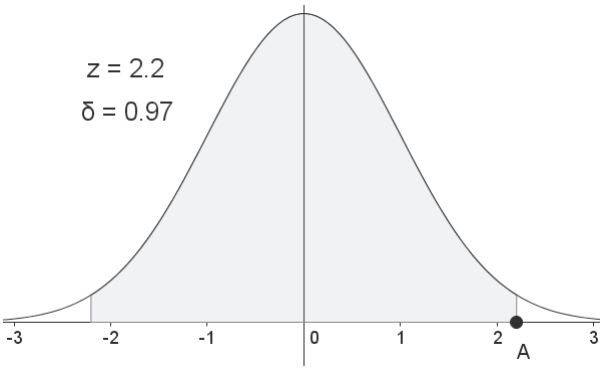

Abb. 4: „Umkehrung“ der Normalverteilung

Direkt kommt man mit

inversNormal[0, 1, (1 + δ)/2]

von δ zu *z*.

6 Konfidenzintervalle für die Einzelerfolgs-Wahrscheinlichkeit bei der Binomialverteilung

Man hat eine Stichprobe vom Umfang *n* und misst $X = k$ als Anzahl der Erfolge.

Mit welchen Einzelerfolgs-Wahrscheinlichkeiten *p* ist dies verträglich?

Diese Frage ist nur sinnvoll, wenn man eine Sicherheits-Wahrscheinlichkeit δ (und damit z) festgelegt hat.

Zu jedem p gehört eine Verteilung, und man kann prüfen, ob die Verteilung mit dem Messergebnis kompatibel ist oder nicht. „Kompatibel“ heißt: Das Messergebnis h muss im zentralen $\delta \cdot 100\%$-Intervall liegen.

Am besten realisiert man das Histogramm einer Binomialverteilung, deren Parameter p über einen Schieberegler variierbar ist (Abb. 5); dabei ist n fest.

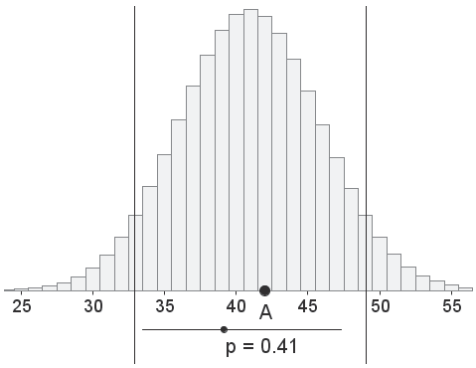

Abb. 5: Zum Konfidenzintervall

Der dicke (variable) Punkt auf der Rechtsachse legt das Messergebnis k fest (in Abb. 5 ist $k = 42$ sowie $n = 100$ sowie $z = 1,64$). Man bekommt die beiden senkrechten Striche einfach durch

x=μ-z*σ und x=μ+z*σ.

Variiert man nun den Parameter p, so sieht man, dass das Messergebnis k genau dann zwischen den beiden senkrechten Strichen (und damit im zentralen $\delta \cdot 100\%$-Intervall) liegt, wenn die Ungleichung

$$0,34 \le p \le 0,5$$

erfüllt ist. Damit hat man das zu k gehörige Konfidenzintervall gewonnen.

7 Konfidenzellipse für die Einzelerfolgs-Wahrscheinlichkeit bei der Binomialverteilung

Die zu k gehörige relative Häufigkeit ist $h = \frac{k}{n}$.

Man hat die Doppel-Ungleichung

$$\underbrace{p - z \cdot \sqrt{\frac{p \cdot (1-p)}{n}}}_{u(p)} \le h \le \underbrace{p + z \cdot \sqrt{\frac{p \cdot (1-p)}{n}}}_{v(p)};$$

die linke und die rechte Seite sind jeweils von p abhängige Funktionsterme $u(p)$ und $v(p)$. Das legt es nahe, sich die zugehörigen Graphen anzusehen; h muss dann zwischen den Funktionswerten zu u (unten) und zu v (oben) liegen (Abb. 6; hier wurde $n = 100$ und $\delta = 0,9$ gewählt). Leider kann man die Rechtsachse (noch) nicht konsequent in p-Achse umbenennen; man muss also

u(x)=x-z*sqrt(x*(1-x)/n)

eingeben; analog v(x).

Zu h gehört der Punkt H auf der Hochachse (am einfachsten definiert man H als variablen Punkt auf der Hochachse). Dann liegt h = y(H) genau dann zwischen $u(p)$ und $v(p)$, wenn (mit dem Wert $h = 0,42$ von Abb. 5) die Ungleichung

$$0,34 \le p \le 0,5$$

erfüllt ist. Dies ist ein anderer Zugang zum zu h gehörigen Konfidenzintervall.

Die Graphen zu u und v bilden zusammen eine Ellipse, die Konfidenzellipse. (Dass es sich tatsächlich um eine Ellipse handelt, kann in diesen kegelschnittfernen Zeiten im Unterricht kein Thema werden.)

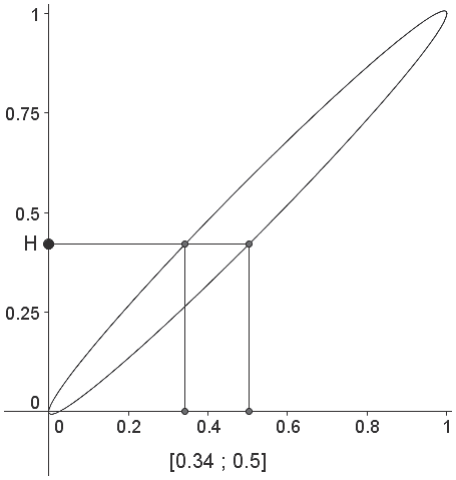

Abb. 6: Konfidenzintervall und Konfidenzellipse

8 Zur Überdeckungs-Wahrscheinlichkeit bei der Binomialverteilung

Da die Messgröße k (bzw. h) eine Zufallsvariable ist, sind auch die Grenzen des zugehörigen Konfidenzintervalls Zufallsvariablen. Es kann auch sein, dass das ermittelte Konfidenzintervall den wahren (aber leider unbekannten) Wert p gar nicht überdeckt. Man wird fragen, mit welcher Wahrscheinlichkeit das der Fall ist.

Nun hilft hypothetisches Denken: Nehmen wir an, der wahre (aber leider unbekannte) Wert sei *w*. Dann ist *k* (bzw. *h*) Wert einer binomialverteilten Zufallsgröße *X* (mit Parametern *n* und *w*). Mit Wahrscheinlichkeit δ gilt:

$$\mu - z \cdot \sigma \le k \le \mu + z \cdot \sigma$$

bzw. (nach Division durch *n*)

$$w - z \cdot \sqrt{\frac{w \cdot (1-w)}{n}} \le h \le w + z \cdot \sqrt{\frac{w \cdot (1-w)}{n}} .$$

In Abb. 7 ist diese *w*-Umgebung auf der Hochachse (für $w = 0{,}4$) sichtbar gemacht.

Man sieht ohne (!) zugehörige Rechnung: Genau dann, wenn *h* außerhalb der erwähnten *w*-Umgebung liegt, überdeckt das zu *h* gehörige Konfidenzintervall den wahren Wert *w* nicht. In Abb. 7 wurde $h = 0{,}6$ zur Illustration gewählt; das zugehörige Konfidenzintervall überdeckt *w* nicht. (Die beiden Achsen müssen hier gleich skaliert sein!)

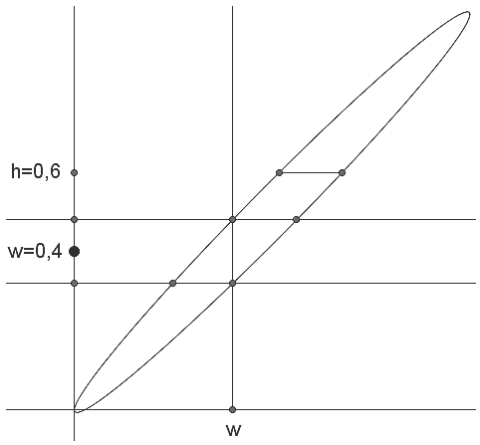

Abb. 7: Zur Überdeckungs-Wahrscheinlichkeit bei der Binomialverteilung

Überdeckungs-Wahrscheinlichkeit und Sicherheits-Wahrscheinlichkeit haben also denselben Wert.

Mit der Konfidenzellipse kann man gut die bei Abänderung von δ oder von *n* eintretenden Effekte anschaulich machen, da sowohl δ als auch *n* durch Schieberegler festgelegt werden und damit variabel sind.

9 Die Grenzen der Konfidenzintervalle sind Zufallsgrößen

Wenn *w* der wahre Wert der Einzelerfolgs-Wahrscheinlichkeit ist, ist die absolute Häufigkeit $n \cdot h$ eine (*n*, *w*)-binomialverteilte Zufallsgröße; jede Messung liefert ein anderes *h* und damit ein anderes Konfidenzintervall. Dies lässt sich in GeoGebra anschaulich machen:

Mit h=ZufallszahlBinomialverteilt[n, w]/n wird *h* erzeugt. Nun kann man den Punkt *H* in Abb. 6 als (0, h) definieren. Mit F9 bekommt man eine neue Zufallszahl *h* und damit einen neuen Punkt *H*.

Noch überzeugender ist es, wenn man in einer einzigen Graphik 10-mal *h* und das zugehörige Konfidenzintervall erzeugen und anzeigen lässt. (Auch hier ist eine Demonstration seitens der Lehrperson völlig ausreichend; das folgende Verfahren sollten Schülerinnen und Schüler nicht lernen müssen.)

Löst man die grundlegende Doppel-Ungleichung

$$p - z \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \le h \le p + z \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

nach *p* auf, so bekommt man die Lösungen $p = c \pm \sqrt{c^2 - d}$ der zugehörigen quadratischen Gleichung als Grenzen des Konfidenzintervalls; dabei sind $c = \frac{n \cdot h + 0{,}5 \cdot z^2}{n + z^2}$ und $d = \frac{n \cdot h^2}{n + z^2}$.

Nach Definition von *n*, δ und *w* werden mit

```
Lh=Folge[ZufallszahlBinomialverteilt[n, w] / n,
         k, 1, 10]
```

zehn Zufallszahlen der gewünschten Art erzeugt.

Mit

```
Lc=(n*Lh+0.5*z^2)/(n+z^2)
```

und

```
Ld=n*Lh^2/(n+z^2)
```

bekommt man die Listen der zugehörigen Werte *c* und *d*.

Die Grenzen der Konfidenzintervalle werden durch

```
Llinks=Lc-sqrt(Lc^2-Ld)
```

und durch

```
Lrechts=Lc+sqrt(Lc^2-Ld)
```

geliefert; die zugehörigen Punkte erhält man durch

```
Plinks=Folge[(Element[Llinks, k], k), k, 1, 10]
```

und

```
Prechts=Folge[(Element[Lrechts, k], k), k, 1, 10].
```

Nun braucht man nur noch die Strecken zwischen den Endpunkten:

```
Folge[Strecke[Element[Plinks, k],
              Element[Prechts, k]], k, 1, 10].
```

Wenn man will, kann man sich mit

Ph= Folge[(Element[Lh, k], k), k, 1, 10]

zu jedem Konfidenzintervall auch den zugehörigen *h*-Wert anzeigen lassen.

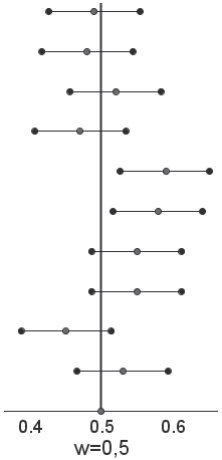

Abb. 8: Mehrere Konfidenzintervalle zu w = 0,5

In Abb. 8 ist $n = 100$, $\delta = 0{,}8$ und $w = 0{,}5$. Erwartungsgemäß überdecken nur 80 % der Konfidenzintervalle den wahren Werte.

Mit F9 bekommt man zehn neue Konfidenzintervalle. Natürlich erreicht man nicht immer eine Überdeckungsrate von 80 %.

Anschrift des Verfassers

Dr. Jörg Meyer
Schäfertrift 16
31789 Hameln
J.M.Meyer@t-online.de