

Konfidenzintervalle I

Situation: Binomialverteilung und Sigmaregeln sind bekannt. Die Sigmaregeln wurden ohne Hinweis auf die Normalverteilung eingeführt. Der Begriff der Normalverteilung soll in dieser Version auch weiterhin nicht explizit verwendet werden. Eine solche Version des Unterrichtsganges scheint - zumindest als benchmark - notwendig, da die Normalverteilung NICHT explizit behandelt werden muss.

Für eine Gesundheitsstudie werden 100 Mannheimer auf ihre Blutgruppe untersucht. Wir interessieren uns für die unbekannte Wahrscheinlichkeit p , dass ein Mannheimer die Blutgruppe 0 hat. Die Untersuchung liefert 40 „Treffer“, d.h. 40 Mannheimer aus dieser Stichprobe haben die Blutgruppe 0. Da die 100 Mannheimer zufällig ausgewählt wurden, genügt es aber nicht, die unbekannte Wahrscheinlichkeit mittels der ermittelten relativen Häufigkeit zu schätzen.

Wir suchen daher ein Intervall, das die unbekannte (Erfolgs)wahrscheinlichkeit p enthält. Ein solches Intervall nennt man **Vertrauensintervall**.

Nach der Sigmaregel liegt die beobachtete Trefferzahl k mit 95% Wahrscheinlichkeit im Intervall $[\mu - 1,96\sigma; \mu + 1,96\sigma]$. Daher gilt mit einer Wahrscheinlichkeit von 95%

$$n \cdot p - 1,96\sqrt{n \cdot p(1-p)} \leq k \leq n \cdot p + 1,96\sqrt{n \cdot p(1-p)}$$

bzw. für beliebige Sicherheitswahrscheinlichkeiten

$$n \cdot p - z\sqrt{n \cdot p(1-p)} \leq k \leq n \cdot p + z\sqrt{n \cdot p(1-p)}.$$

Teilt man durch den Stichprobenumfang n , erhält man für die relative Häufigkeit $h = \frac{k}{n}$

$$p - z\sqrt{\frac{p(1-p)}{n}} \leq h \leq p + z\sqrt{\frac{p(1-p)}{n}} \quad (1).$$

Gesucht sind also alle Werte von p , die diese Ungleichung erfüllen. Im folgenden sieht man nun, wie man diese Werte für p graphisch und rechnerisch finden kann. In beiden Fällen jedoch wird eine Näherungslösung verwendet. Die Lösungsformel für die exakten Werte wird ebenfalls angesprochen, dient jedoch lediglich als Ausgangspunkt für die rechnerische Lösung.

Arbeitsauftrag und Hinweise (kursiv): Die Arbeitsaufträge sind in dieser Form eher für Lehrer als für Schüler gedacht, können aber leicht für Schüler umgeschrieben werden.

1. Wir bestimmen nun - zunächst graphisch - eine Näherung für das Vertrauensintervall.

Öffnen Sie dazu die Geogebra-Datei *Vertrauen1.ggb*

Bestimmen Sie - durch Veränderung von p - die Grenzen des 95%-Vertrauensintervalls.

Man verändere p so lange, bis der Punkt A mittig zwischen den beiden Strichen liegt. Das hier ist also eine Möglichkeit, das Problem graphisch zu lösen. Natürlich kann und sollte man diesen Teil des Arbeitsauftrages noch erweitern und weitere Intervalle.

2. Zeigen Sie, dass sich aus Gleichung (1) die exakten Grenzen des Vertrauensintervalls ergeben:

$$p = \frac{h + \frac{z^2}{2n} \pm \frac{z}{\sqrt{n}} \sqrt{h(1-h) + \frac{z^2}{4n}}}{1 + \frac{z^2}{n}} \quad (*)$$

Wenn man alle Terme mit $\frac{z^2}{n}$ vernachlässigt (aber nicht den Term mit $\frac{z}{\sqrt{n}}$), ergibt sich die rechnerische Näherungslösung für die Intervallgrenzen:

$$p = h \pm z \sqrt{\frac{h(1-h)}{n}}$$

Berechnen Sie nun das 95%-Vertrauensintervall mit Hilfe dieser Näherungsformel .

Die graphische Lösung von oben war nicht die exakte Lösung.

Terme in () wegstreichen, (sich) klar machen, dass das geht, wenn „n sehr gross“ im Vergleich zu z ist, Grenzen nun erneut mittels der Formel berechnen. Man bekommt die rechnerische Näherungslösung.*

Dieser Weg zur Formel benutzt also wenigstens an einer Stelle explizit die exakte Lösungsformel. Geht man diesen Weg nicht, muss man den Schülern irgendwie anders klar machen, warum aus Gleichung (1) die Näherungsformel hervorgeht. Die Formel und Gleichung (1) unterscheiden sich ja sozusagen nur dadurch, dass sich der Term unter der Wurzel ändert und man sich das Intervall quasi (irgendwie) zusammenbastelt. Es gibt sicher Schüler, die diese Vorgehensweise etwas merkwürdig finden.

Die graphische Ermittlung („Hyperbel“) der exakten Lösung findet sich in der Datei 51BspFII.ggb.

(siehe auch Lambacher Schweizer, Stochastik, Kap. 5.)

3. Öffnen Sie die Geogebra-Datei *Vertrauen2.ggb*

Wir nehmen nun einen fiktiven, „wahren“ Wert für p an und simulieren 100 Messungen in 100 Krankenhäusern, jede Messung liefert ein anderes Vertrauensintervall, zunächst für die Sicherheitswahrscheinlichkeit $\delta = 0.95$. Schildern Sie Ihre Beobachtungen über

- die Symmetrie der Intervallgrenzen zum wahren p
- die Anzahl der Vertrauensintervalle, die das wahre p überdecken
- Länge des Intervalls für großes bzw. kleines n
- die Werte für die Intervallgrenzen für kleine Werte von n bzw. p

Mit F9 starten Sie eine erneute Simulation von 100 Messungen.

Wie erklären sich die jeweiligen Beobachtungen?

Lage: die Intervalle liegen „irgendwo“ und müssen insbesondere nicht symmetrisch zum wahren p liegen. Warum? Nun, die Intervallgrenzen sind Zufallsvariablen und bei jeder Untersuchung kann was anderes herauskommen. Damit ist auch klar, dass die oft gehörte Aussage: „ p liegt mit W 'keit von 95% im Intervall mit den Grenzen a und b “ falsch ist. Nicht p ist die Zufallsgröße, sondern die Intervallgrenzen. Also: Wenn wir dieselbe Untersuchung sehr oft wiederholten, dann würde das tatsächliche p mit 95% Wahrscheinlichkeit der Fälle in dem jeweils von uns ermittelten Intervall liegen. Dieses Intervall ist aber für jede Untersuchung ein anderes, da es aus den jeweiligen Beobachtungswerten berechnet wird.

Länge des Intervalls: je mehr Beobachtungen ich habe, desto „sicherer“ kann ich mir sein, dass der wahre Parameter von den Intervallen überdeckt wird. Also werden die Intervalle kleiner.

Zahl: nicht alle Intervalle überdecken das wahre p überhaupt. Aber es sollte immer eine Zahl sein, die sehr nahe an der Sicherheitswahrscheinlichkeit von 95% liegt.

Werte der Grenzen: für kleine Werte von n und p kann es sein, dass die Intervallgrenzen negativ bzw. größer als 1 werden. Das liegt daran, dass wir eben nicht die exakten Intervallgrenzen berechnen, sondern eine Näherung verwenden. Insbesondere bricht die Normalverteilung nicht für große bzw. kleine Werte ab, sondern nähert sich asymptotisch der x -Achse, was bei der Binomialverteilung nicht der Fall ist.

Die Näherungslösung macht noch andere Probleme, insbesondere sind die Überdeckungswahrscheinlichkeiten nicht nur für kleine Werte von n und/oder p oft unbefriedigend. Mehr dazu bei der Station „Wissenschaftliche Artikel zu Konfidenzintervallen.“