Konfidenzintervall für den Parameter p einer Binomialverteilung $B(n,p)$

Ausgangssituation

Man ziehe aus einer (unendlich groß gedachten) Population eine Stichprobe des Umfangs n. Die Personen werden nach ihrer Einstellung zu einem bestimmten Thema gefragt und haben genau 2 Antwortmöglichkeiten: $+$ oder $-$.

k sei die Anzahl der „+“-Antworten in der Stichprobe (absolute Häufigkeit).

Die entsprechende relative Häufigkeit ist $r = \frac{k}{n}$.

p sei der wahre Anteil der „+“-Stimmen in der Gesamtpopulation.

Die Verteilungen

K sei die Zufallsvariable, die durch wiederholte Ziehung unabhängiger Stichproben und Berechnung von k entsteht. Analog zur einzelnen Stichprobe gilt $R = \frac{K}{n}$.

K ist binomialverteilt mit $E(K) = \mu_K = n \cdot p$ und $\sigma_K = \sqrt{n \cdot p \cdot q}$.

Die Binomialverteilung $B(n,p)$ darf durch die Normalverteilung $N(\mu_K, \sigma_K)$ ersetzt werden, falls $n \cdot p \geq 5$ und $n \cdot q \geq 5$. Da p und q unbekannt sind, muss hier stattdessen (als Schätzung) $n \cdot r$ und $n \cdot (1-r)$ verwendet werden. Da r ein erwartungstreuer und konsistenter Schätzer für p ist, ist anzunehmen, dass $r \approx p$ und $1 - r \approx q$ gilt. Sind also $n \cdot r$ und $n \cdot (1-r)$ deutlich größer als 5, so kann man getrost die NV-Approximation durchführen:

$$K \sim B(n,p) \rightarrow K \overset{a.}{\sim} N(\mu_K, \sigma_K)$$

Für die Verteilung von $R = \frac{K}{n}$ folgt:

$$R \overset{a.}{\sim} N(\frac{\mu_K}{n}, \frac{\sigma_K}{n}) = N(p, \sqrt{\frac{pq}{n}})$$

Standardisierung...

$$\frac{R - E(R)}{\sigma_R} = \frac{R - p}{\sqrt{pq/n}} \sim N(0,1)$$

Wir wissen, dass für 95% der Werte jeder standardnormalverteilten Variable gilt:

$$|Z| \leq 1.96$$

also auch:

$$\left| \frac{R - p}{\sqrt{pq/n}} \right| = \frac{|R - p|}{\sqrt{pq/n}} \leq 1.96$$

$$|R - p| \leq 1.96 \cdot \sqrt{\frac{pq}{n}}$$

Die (unbekannte!) Populationsvarianz wird nun durch eine (bias-korrigierte) Schätzung ersetzt:

$$\sigma_R = \sqrt{\frac{pq}{n}} \longrightarrow \hat{\sigma}_R = \sqrt{\frac{r(1-r)}{n-1}}$$

Somit gilt:

$$|R - p| \leq 1.96 \cdot \sqrt{\frac{r(1-r)}{n-1}}$$

Die Grenzen des 95%-KI

$$p_{1,2} = r \pm 1.96 \cdot \sqrt{\frac{r(1-r)}{n-1}}$$

Allgemein: Die Grenzen für ein KI mit Irrtumswkt. α

$$p_{1,2} = r \pm z_{1-\alpha/2} \cdot \sqrt{\frac{r(1-r)}{n-1}}$$

Beispiel

Am Abend der Präsidentschafts-Wahl werden $n = 300$ Personen befragt, ob sie zur Wahl gegangen sind. $k = 267$ antworten mit „Ja“. Eine erste Schätzung des Anteils in der Gesamtbevölkerung ist demnach $\hat{p} = r = \frac{k}{n} = 0.89$.

Berechnen des KI:

$$p_{1,2} = 0.89 \pm 1.96 \cdot \sqrt{\frac{0.89 \cdot 0.11}{299}}$$

$$p_1 \approx 0.8545$$

$$p_2 \approx 0.9255$$

$$KI : [0.8545; 0.9255]$$

Der „wahre“ Anteil der Personen in der Bevölkerung, die zur Wahl gegangen sind, liegt (mit einer Wahrscheinlichkeit von 95%) zwischen 85.45% und 92.55%.