Schluss von der Stichprobe auf die Gesamtheit

Problem: Auf Grund einer Stichprobe soll die Wahrscheinlichkeit p einer Binomialverteilung mit einer vorgegebenen Irrtumswahrscheinlichkeit ermittelt werden.

Beispiel: Vor einer Wahl möchte die Partei C gerne wissen, ob sie in ihrem Bundesland die absolute Mehrheit erringen wird. Es werden n=1000 Personen befragt und H_{1000} =538 davon geben an, Partei C wählen zu wollen. Kann die Partei mit 95%-iger Sicherheit davon ausgehen, die absolute Mehrheit zu erringen?
Die Frage soll entschieden werden, indem alle Ergebnisse (also alle p) berechnet werden, für die der Umfragewert 538 innerhalb der 95%-Umgebung liegt.

Für eine 95%-Umgebung gilt die Ungleichung $\mu-1{,}96\cdot\sigma\leq H_n\leq\mu+1{,}96\cdot\sigma$ oder ausführlicher
$n\cdot p-1{,}96\cdot\sqrt{n\cdot p\cdot(1-p)}\leq H_n\leq n\cdot p+1{,}96\cdot\sqrt{n\cdot p\cdot(1-p)}$

Angewendet auf das Beispiel ergibt sich die Ungleichung
$1000\cdot p-1{,}96\cdot\sqrt{1000\cdot p\cdot(1-p)}\leq 538\leq 1000\cdot p+1{,}96\cdot\sqrt{1000\cdot p\cdot(1-p)}$

Im Extremfall (also für die Randwerte des Intervalls der gesuchten p-Werte) gelten die beiden Gleichungen

(1) $1000\cdot p-1{,}96\cdot\sqrt{1000\cdot p\cdot(1-p)}=538$ und (2) $538=1000\cdot p+1{,}96\cdot\sqrt{1000\cdot p\cdot(1-p)}$

Aus Gleichung (1) folgt (3) $1000\cdot p-538=1{,}96\cdot\sqrt{1000\cdot p\cdot(1-p)}$

Aus Gleichung (2) folgt (4) $538-1000\cdot p=1{,}96\cdot\sqrt{1000\cdot p\cdot(1-p)}$

Um die Wurzeln zu entfernen, werden die Gleichungen (3) und (4) quadriert:

Gleichung (3): $(1000\cdot p-538)^2=1{,}96^2\cdot(1000\cdot p\cdot(1-p))\ \rightarrow$
$1000000\cdot p^2-2\cdot 1000\cdot p\cdot 538+538^2=1{,}96^2\cdot 1000\cdot p-1{,}96^2\cdot 1000\cdot p^2$ (5)

Gleichung (4): $(538-1000\cdot p)^2=1{,}96^2\cdot(1000\cdot p\cdot(1-p))\ \rightarrow$
$538^2-2\cdot 1000\cdot p\cdot 538+1000000^2\cdot p^2=1{,}96^2\cdot 1000\cdot p-1{,}96^2\cdot 1000\cdot p^2$ (6)

Da die Gleichungen (5) und (6) übereinstimmen, wird für beide zusammen weiter umgeformt:

$$1000000\cdot p^2-1076000\cdot p+289444=3841{,}6\cdot p-3841{,}6\cdot p^2\ \rightarrow$$

$$1003841{,}6\cdot p^2-1079841{,}6\cdot p+289444=0\ \rightarrow\ p^2-\frac{1079841{,}6}{1003841{,}6}\cdot p+\frac{289444}{1003841{,}6}=0\ \rightarrow$$

$$p^2-\frac{10798416}{10038416}\cdot p+\frac{2894440}{10038416}=0\ \rightarrow\ p_{1,2}=+\frac{5399208}{10038416}\pm\sqrt{\frac{5399208}{10038416}^2-\frac{2894440}{10038416}}\ \rightarrow$$

$p_1=0{,}507$; $p_2=0{,}569$

Mit 95%-iger Sicherheit wird also die Partei C ein Wahlergebnis zwischen 50,7% und 56,9% erringen und damit die absolute Mehrheit erreichen.

Allgemeine Rechnung: (mit c als Maß für die Irrtumswahrscheinlichkeit)

$n\cdot p-c\cdot\sqrt{n\cdot p\cdot(1-p)}\leq H_n\leq n\cdot p+c\cdot\sqrt{n\cdot p\cdot(1-p)}$

Ränder des p-Intervalls: $n\cdot p-c\cdot\sqrt{n\cdot p\cdot(1-p)}=H_n$ und $H_n=n\cdot p+c\cdot\sqrt{n\cdot p\cdot(1-p)}$

Daraus folgt $n\cdot p-H_n=c\cdot\sqrt{n\cdot p\cdot(1-p)}$ und $H_n-n\cdot p=c\cdot\sqrt{n\cdot p\cdot(1-p)}$

Quadrieren ergibt in beiden Gleichungen $n^2 \cdot p^2 - 2 \cdot n \cdot p \cdot H_n + H_n^2 = c^2 \cdot n \cdot p - c^2 \cdot n \cdot p^2 \;\rightarrow$

$$(n^2 + c^2 \cdot n) \cdot p^2 - (2 \cdot n \cdot H_n + c^2 \cdot n) \cdot p + H_n^2 = 0 \;\rightarrow\; p^2 - \frac{2 \cdot n \cdot H_n + c^2 \cdot n}{n^2 + c^2 \cdot n} \cdot p + \frac{H_n^2}{n^2 + c^2 \cdot n} = 0 \;\rightarrow$$

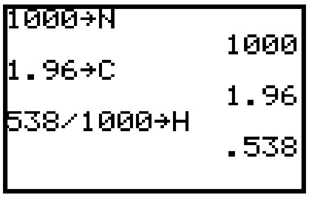
$$p^2 - \frac{2 \cdot H_n + c^2}{n + c^2} \cdot p + \frac{H_n^2}{n \cdot (n + c^2)} = 0 \;\rightarrow\; p_{1,2} = \frac{2 \cdot H_n + c^2}{2 \cdot (n + c^2)} \pm \sqrt{\frac{(2 \cdot H_n + c^2)^2}{4 \cdot (n + c^2)^2} - \frac{H_n^2}{n \cdot (n + c^2)}}$$

Die allgemeine Formel für p ist etwas unhandlich. Wenn die solver-Funktion des Taschenrechners benutzt werden darf, sollte man eine einfachere Gleichung vorziehen, z.B. $|n \cdot p - H_n| = c \cdot \sqrt{n \cdot p \cdot (1-p)}$, die sich aus $n \cdot p - H_n = c \cdot \sqrt{n \cdot p \cdot (1-p)}$ und $H_n - n \cdot p = c \cdot \sqrt{n \cdot p \cdot (1-p)}$ ergibt.

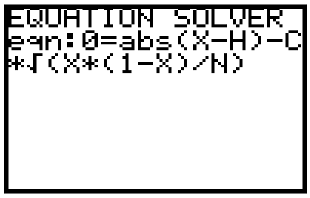
Setzt man dann noch statt der absoluten Häufigkeit H_n die relative Häufigkeit $h = \frac{H_n}{n}$ ein, so ergibt sich $|n \cdot p - n \cdot h| = c \cdot \sqrt{n \cdot p \cdot (1-p)} \;\rightarrow\; |p - h| = \frac{c \cdot \sqrt{n \cdot p \cdot (1-p)}}{n} = c \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$.

Die Lösung des Beispiels mit dem Taschenrechner ergibt sich so:

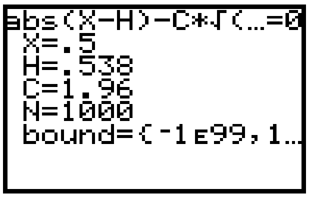
Konstanten-Werte eingeben

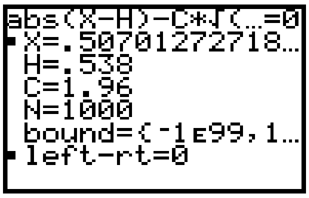


Gleichung eingeben

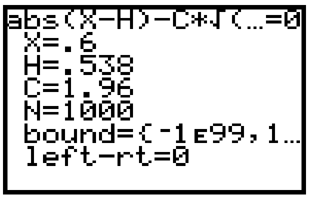


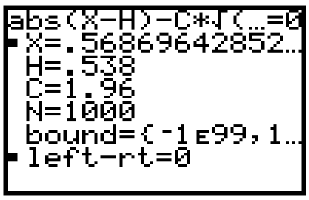
Näherung p=0,5 gibt 0,507





Näherung p=0,6 gibt 0,569





Ist die Intervalllänge d des Vertrauensintervalls vorgegeben und wird nach dem für diese Intervalllänge notwendigen Stichprobenumfang gefragt, so ergibt sich aus $|p - h| = c \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$ für die Intervalllänge $2 \cdot |p - h| = 2 \cdot c \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \le d$

Der Wert für $p \cdot (1-p)$ wird maximal für $p = \frac{1}{2}$. Mit diesem Wert muss die Intervalllänge noch kleiner oder gleich d sein: $2 \cdot c \cdot \sqrt{\frac{\frac{1}{2} \cdot \left(1 - \frac{1}{2}\right)}{n}} \le d \;\rightarrow\; 4 \cdot c^2 \cdot \frac{\frac{1}{4}}{n} \le d^2 \;\rightarrow\; n \ge \frac{c^2}{d^2}$