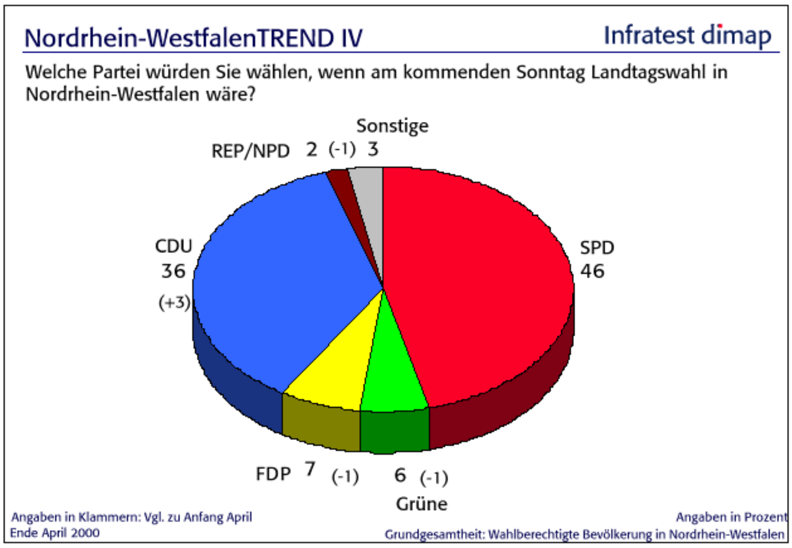
Konfidenzintervalle

Im April 2000 führte Infratest dimap unter 1000 Wahlberechtigten in NRW eine Umfrage zur bevorstehenden Landtagswahl durch. Bei den angegebenen Ergebnissen der Umfrage handelt es sich natürlich um relative Häufigkeiten. Für die SPD wurde dabei eine relative Häufigkeit von h=46% ermittelt. Die absolute Häufigkeit der SPD-Wähler betrug also x=460. Die Wahrscheinlichkeit p dafür, dass ein Wahlberechtigter die SPD wählt, ist natürlich nicht genau 46%. Sie liegt lediglich nahe bei 46%. Doch was heißt nun *"nahe bei"*?



Ein Wert p für die tatsächliche Wahrscheinlichkeit kommt dann in Frage, wenn der Erwartungswert der entsprechenden Binomialverteilung nahe bei der in der Umfrage mit n=1000 Befragten ermittelten absoluten Häufigkeit x=460 liegt.
Die Zufallsgröße X beschreibt dann die Anzahl der SPD-Wähler. Sie ist B-1000,p - verteilt.

Aufgabe 1

Zeichne für verschiedene Werte von p ein Balkendiagramm der Binomialverteilung.
Begründe die folgende Entscheidungsregel dafür, dass der zugrundeliegende Wert von p akzeptiert oder abgelehnt wird.

Entscheidungsregel: Falls x=460 innerhalb eines Bereiches um den Erwartungswert mit der Wahrscheinlichkeit von z.B. 95% liegt, wird die zugrundeliegende Wahrscheinlichkeit akzeptiert, anderenfalls wird sie abgelehnt. (Bei vorsichtigerer Vorgehensweise können wir natürlich statt eines 95%-Bereiches auch einen 99%-Bereich vorsehen.)

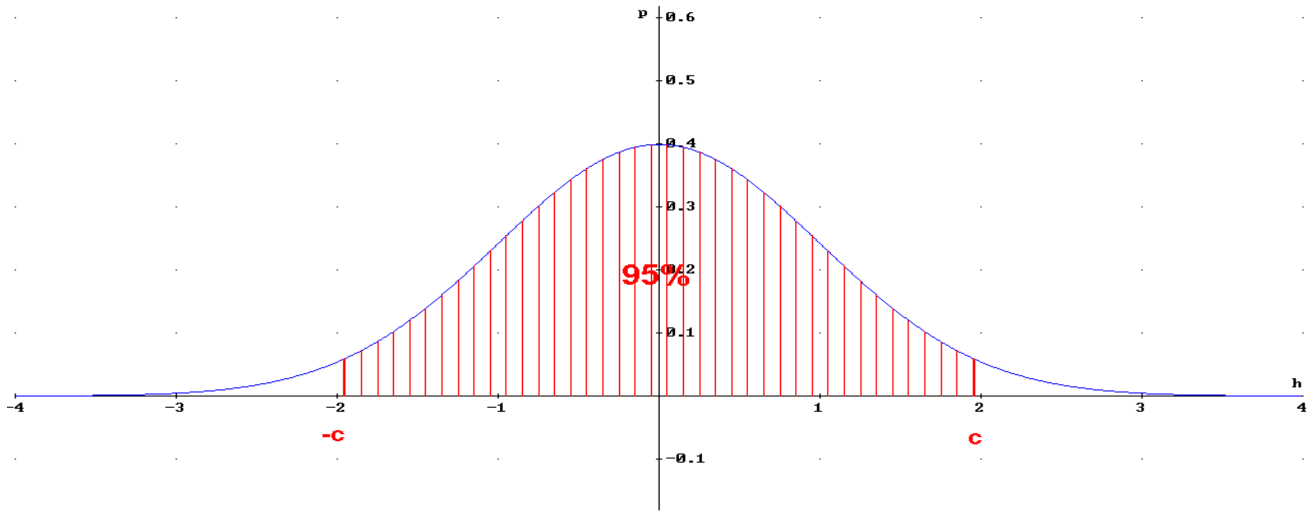
Aufgabe 2

Schätze ein Intervall [p1, p2] ab für die Wahrscheinlichkeit, einen SPD-Wähler zu treffen.

Aufgabe 3

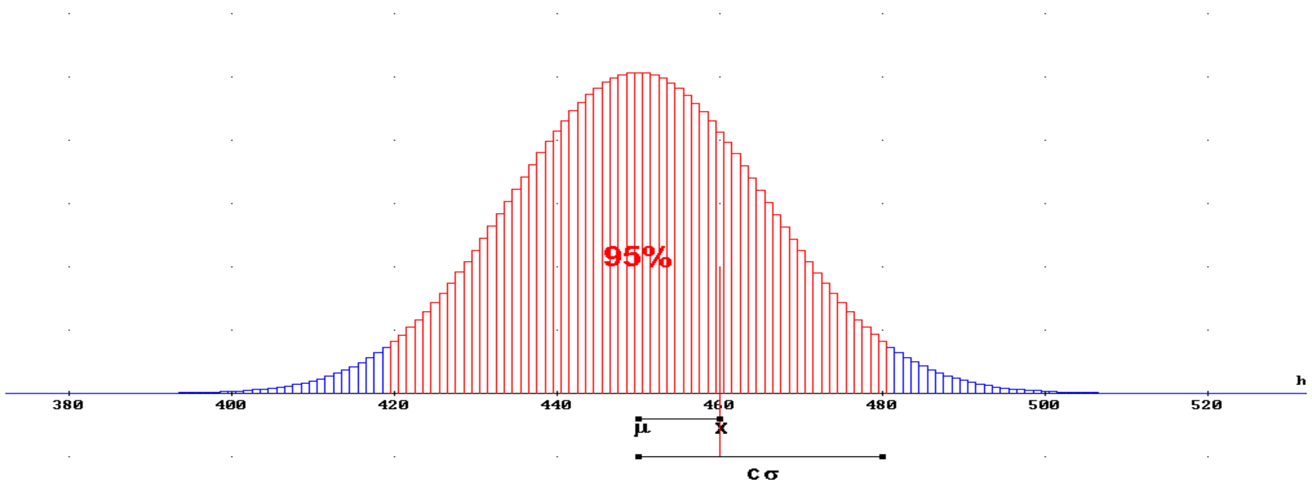
Das in der letzten Aufgabe näherungsweise bestimmte Intervall bezeichnen wir als das **Vertrauensintervall** für die unbekannte Wahrscheinlichkeit p oder auch als **Konfidenzintervall**. Die vorgegebene Wahrscheinlichkeit des Bereiches heißt **Vertrauenszahl γ**. Hier ist also γ=0.95.

In dieser Aufgabe suchen wir nach einer Möglichkeit, das Vertrauensintervall rechnerisch ohne Verwendung des vector-Befehls zu ermitteln.
Zur Bestimmung der Grenzen des γ-Bereichs betrachten wir statt der Binomialverteilung selbst zunächst die normalisierte Binomialverteilung.



a) Zeige: $\Phi(c) - \Phi(-c) = 2\Phi(c) - 1$
b) Bestimme c mit $2\Phi(c) - 1 = 0.95$

Wenn c die halbe Breite des 95%-Bereichs der normalisierten Binomialverteilung ist, so beträgt die halbe Breite der ursprünglichen Binomialverteilung $c\sigma$.



Liegt die in der Umfrage ermittelte absolute Häufigkeit x=460 innerhalb des 95%-Bereichs, so gilt: $|x - \mu| \leq c\sigma$
Die Division dieser Ungleichung durch n ergibt: $|h - p| \leq c\sigma/n$

Das gesuchte Vertrauensintervall ist also: **$[h - c\sigma/n\ ,\ h + c\sigma/n]$ mit $\Phi(c) = (1 + \gamma)/2$**

c) Durch Quadrieren der letzten Ungleichung erhält man $(h - p)^2 \leq (c\sigma/n)^2$. Löse diese Ungleichung mit den Werten des Eingangsbeispiels und gib damit das Vertrauensintervall zur Vertrauenszahl γ=0.95 an.

d) Zeige: $P(|h - p| \leq c\sigma/n) = \gamma$

e) Bestimme auch die Vertrauensintervalle zur Vertrauenszahl γ=0.95 für die Wahrscheinlichkeit einen Wähler der CDU, der FDP oder der Grünen anzutreffen.

Aufgabe 4

Mit zunehmendem Stichprobenumfang wird das Vertrauensintervall schmaler. Es ist sinnvoll, vor der Durchführung einer Umfrage den erforderlichen Stichprobenumfang für eine gewünschte Maximalbreite des Vertrauensintervalls abzuschätzen.

a) Zeige: $p(1-p) \leq 1/4$ für alle $p[0\ ,\ 1]$

Für die Breite des Vertrauensintervalls gilt damit:

$$2\frac{c\sigma}{n} = 2c\frac{\sqrt{np(1-p))}}{n} = 2c\sqrt{\frac{p(1-p)}{n}} \leq 2c\sqrt{\frac{1}{4n}} = \frac{c}{\sqrt{n}}$$

Soll diese Intervallbreite kleiner oder gleich der vorgegebenen Breite d sein, so gilt:

$$n \geq \frac{c^2}{d^2}$$

b) Berechne den mindestens erforderlichen Stichprobenumfang für eine maximale Breite des Vertrauensintervalls von d=2.5%.

Aufgabe 5

Mit der Ungleichung in Aufgabe 2 für das Vertrauensintervall sieht man, dass die Breite für größere n abnimmt. Welchen Einfluss hat die relative Häufigkeit auf die Vertrauensintervalle?

Da bei der Untersuchung viele Vertrauensintervalle zu berechnen sind, soll zunächst das Verfahren zur Berechnung automatisiert werden.

a) Löse die Gleichung für die Grenzen des Vertrauensintervalls allgemein.

#1: $$p^2 \cdot \left(1 + \frac{c^2}{n}\right) - p \cdot \left(2 \cdot h + \frac{c^2}{n}\right) + h^2 = 0$$

b) Zur Bestimmung von c ist eine Gleichung der Form $\Phi(x) = u$ zu lösen. Da eine Umkehrfunktion der Gaußschen Summenfunktion nicht elementar angegeben werden kann, lösen wir die äquivalente Gleichung $\Phi(x) - u = 0$ numerisch mit dem Newton-Verfahren. Gib die entsprechende Newton-Iteration mit dem Iterate-Befehl für 10 Iterationsschritte an und definiere eine Funktion wert_c(γ) zur Berechnung von c.

c) Gib nun je eine Funktion zur Berechnung der linken und der rechten Grenze des Vertrauensintervalls in Abhängigkeit von h, n und γ an. Damit kann leicht auch eine Funktion zur Ermittlung des Vertrauensintervalls definiert werden.

d) Trage mit Hilfe des vector Befehls die Vertrauensintervalle für n = 10 (100 ; 1000) und γ = 0.9 graphisch über den relativen Häufigkeiten auf.
