Erwartungswerte einer halben Binomialverteilung

Jürgen Dollinger

10. August 2012

1 Einleitung

In einer Klausur, *Formale Methoden der Informatik für Wirtschaftswissenschaftler* gab es eine Aufgabe, in der bei 10 Aussagen angekreuzt werden musste, ob sie richtig oder falsch seien. Dabei gibt jedes richtige Kreuz einen Punkt, jedes falsche Kreuz jedoch einen Punkt Abzug, da man sonst durch raten ja schon die Hälfte der Punkte kriegen konnte. So könnte man theoretisch von 10 Punkte (alles richtig gekreuzt) bis -10 Punkte (alles falsch gekreuzt) kriegen. Allerdings wurden keine negativen Gesamtpunkte vergeben, so dass, wer die Hälfte oder mehr falsch hatte, 0 Punkte bekam. Insgesamt wird also der Erwartungswert auch bei reinem Raten positiv sein. Da die Aufgabe ziemlich schlecht ausgefallen war–der Durchschnitt lag bei 2.6 Punkten– stellte sich die Frage: Sind die Studenten eigentlich besser als der Zufall? Falls ja um wieviel?

Stellen wir uns zunächst die Frage, wieviel Möglichkeiten es gibt k richtige aus n Fragen zu haben. Natürlich besteht die Möglichkeit, nicht alle Fragen zu bearbeiten, das behandeln wir, wie wenn einfach weniger Fragen gestellt worden wären – zumindest solange die Ratewahrscheinlichkeit bei 50% liegt. In unserem Fall ist also primär $n = 10$. Es gibt nur eine Art 10 richtige zu haben: alle müssen richtig sein. Die Zahl der Möglichkeiten, 9 richtige zu haben ist dagegen 10, denn es kann die erste falsch sein, die zweite usw. Ab da wird's interessant. Die Zahl der Möglichkeiten, 8 richtige zu haben ist, so schint es, 10 mal 9 also 90 Möglichkeiten, denn für die zweite falsche Antwort gibt es nur noch 9 Möglichkeiten zur Verteilung. Das ist aber nur die halbe Wahrheit (oder viel mehr die doppelte), denn es ist egal, ob man zuerst die erste Aufgabe falsch macht und dann die zweite oder umgekehrt. Es sind also nur halb so viele also 45 Möglichkeiten 8 richtige zu haben. Im allgemeinen Fall wird die Zahl der Möglichkeiten k richtige aus n auszuwählen durch den Binomialkoeffizienten $\binom{n}{k}$ beschrieben.

2 Binomialkoeffizienten

Die Binomialkoeffizienten $\binom{n}{k}$ (gesprochen: n über k, oder k aus n, im Englischen auch n choose k) geben die Zahl der Möglichkeiten an, k aus n auszuwählen.

Namensgebend für die Binomialkoeffizienten ist die Binomische Formel. Wer drei binomische Formeln kennt, weiß wahrscheinlich nichts von **der** binomischen Formel:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}. \tag{1}$$

Multipliziert man $(a+b)^n$ ganz klassisch aus, so tritt jedes Produkt mit n Faktoren $abaaabbbab, bbbaaabaaa, \ldots$ genau einmal auf. Fasst man jedoch solche mit gleicher Anzahl a's zusammen, wie in obiger Summe, stellt man fest, dass solche mit k a's eben $\binom{n}{k}$ mal auftreten, ganz ähnlich wie man k richtige Antworten auf n Fragen verteilen kann. So läßt sich die binomische Formel elegant kombinatorisch begründen.

Rekursiv können die Binomialkoeffizienten mit dem Pascalschen Dreieck berechnet werden. Dort steht links ($\binom{n}{0}$) und rechts ($\binom{n}{n}$) eine 1 und dazwischen jeweils die Summe der beiden darüber stehenden Zahlen.

$$\begin{array}{c\|c} n & \binom{n}{k} \\ 1 & 1\ \ 1 \\ 2 & 1\ \ 2\ \ 1 \\ 3 & 1\ \ 3\ \ 3\ \ 1 \\ 4 & 1\ \ 4\ \ 6\ \ 4\ \ 1 \\ 5 & 1\ 5\ 10\ 10\ 5\ 1 \\ & \ldots \end{array} \tag{2}$$

Als Rekursionsformel läßt sich das so formulieren

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \tag{3}$$

mit den Anfangswerten $\binom{n}{0} = \binom{n}{n} = 1$.

3 Berechnung von Erwartungswerten

Wir wissen also, dass es $\binom{n}{k}$ Möglichkeiten gibt, k richtige Antworten aus n zu wählen. Wenn es nun $f(k)$ Punkte für k richtige Antworten gibt, so erhalten

wir die durchschnittliche Punktezahl $\langle f \rangle$, indem wir die Punkte mit der Häufigkeit ihres Auftretens gewichten, aufsummieren und durch die Gesamtzahl aller Möglichkeiten 2^n dividieren:

$$\langle f \rangle = \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} f(k) \tag{4}$$

Beim Ankreuzen ist es nämlich wie bei Binärzahlen: Es gibt 2^n Zahlen mit n Stellen. Außerdem ist auch mit der binomischen Formel $\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k} 1^k 1^{n-k} = (1+1)^n = 2^n$.

4 Binomialverteilung

Etwas anders geschrieben findet sich

$$\langle f \rangle = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n f(k) = \sum_{k=0}^{n} B(k,n)\, f(k) \tag{5}$$

mit $B(k,n) = \binom{n}{k} \left(\frac{1}{2}\right)^n$. Dies lässt sich wiederum verallgemeinern. Wenn man nicht nur rät, also mit der Wahrscheinlichkeit $p = 0.5$ ankreuzt, sondern mit einer anderen Wahrscheinlichkeit p so tritt an die Stelle von $B(k,n)$ die Binomialverteilung [1] $B(k|p,n) = \binom{n}{k} p^k (1-p)^{n-k}$.

Interessant ist es, ein paar Spezialfälle zu betrachten. Für p sehr nahe bei 1 ist $(1-p)^{n-k}$ praktisch Null außer für $n = k$. Damit ist

$$B(k|1,n) = \begin{cases} 0 & n > k \\ 1 & n = k. \end{cases} \tag{6}$$

Es tritt also nur der Fall auf, dass alles richtig ist.

Für $p = 0.5$ ist natürlich auch $1 - p = 0.5$ und wir erhalten den obigen Spezialfall $B(k|\frac{1}{2},n) = \binom{n}{k} \left(\frac{1}{2}\right)^n$.

5 Punktefunktion

An dieser Stelle sollten wir die Punktefunktion $f(k)$ betrachten. Wieviele Punkte bekommt man bei k Richtigen? Zunächst ist klar, dass bei k richtigen Antworten $(n-k)$ falsche dabei sind. Die Punktezahl setzt sich aus den k Punkten für die richtigen und den Abzügen $-(n-k)$ für die $(n-k)$ falschen Antworten zusammen. Es gilt also $f(k) = k - (n-k) = 2k - n$. Allerdings nur falls dies ≥ 0 ist. Also

$$f(k) = \begin{cases} 2k - n & k > \frac{n}{2} \\ 0 & \text{sonst} \end{cases} \tag{7}$$

Aus der Summe (4) wird also die "Halb"–Summe

$$\langle f \rangle = \frac{1}{2^n} \sum_{k=[\frac{n}{2}]+1}^{n} \binom{n}{k} (2k - n) \tag{8}$$

wobei $[\frac{n}{2}]$ die größte ganze Zahl kleiner oder gleich $\frac{n}{2}$ ist. Bei geraden Zahlen kommt es gar nicht so genau darauf an, ob man bei $\frac{n}{2}$ oder $\frac{n}{2}+1$ beginnt, da der Summand für $k = \frac{n}{2}$ sowieso Null ist.

Mit einer allgemeinen Wahrscheinlichkeit p ($> \frac{1}{2}$) wäre das

$$\langle f \rangle = \sum_{k=[\frac{n}{2}]+1}^{n} \binom{n}{k} p^k (1-p)^{n-k} (2k - n) \tag{9}$$

Hier läßt sich der Normierungsfaktor $p^k(1-p)^{n-k}$ nicht mehr als Konstante $\frac{1}{2^n}$ vor die Summe ziehen.

6 Numerische Ergebnisse

Eine Berechnung der Summe (8) und vor allem (9) in geschlossener Form ist schwierig. Wir befassen uns damit in Abschnitt 8.

Betrachten wir zunächst noch einmal den Fall reinen Ratens ($p = 0.5$, Gleichung (8)). Ein C-Programm wie aus Anhang A berechnet uns einige numerische Werte.

Für $n = 10$ ergibt sich:

$$\begin{aligned}
\langle f \rangle &= \frac{1}{1024} \sum_{k=5}^{10} \binom{n}{k} (2k - n) && (10)\\
&= \frac{1}{1024} \left(\binom{10}{5} \cdot 0 + \binom{10}{6} \cdot 2 + \binom{10}{7} \cdot 4 + \binom{10}{8} \cdot 6 \right. && (11)\\
&+ \left. \binom{10}{9} \cdot 8 + \binom{10}{10} \cdot 10 \right) && (12)\\
&= 1260/1024 = 1.230469 && (13)
\end{aligned}$$

Und für $n = 9$

$$\begin{aligned}
\langle f \rangle &= \frac{1}{512} \sum_{k=5}^{9} \binom{n}{k} (2k - n) && (14)\\
&= \frac{1}{512} \left(\binom{9}{5} \cdot 1 + \binom{9}{6} \cdot 3 + \binom{9}{7} \cdot 5 + \binom{9}{8} \cdot 7 + \binom{9}{9} \cdot 9 \right) && (15)\\
&= 630/512 = 1.230469 && (16)
\end{aligned}$$

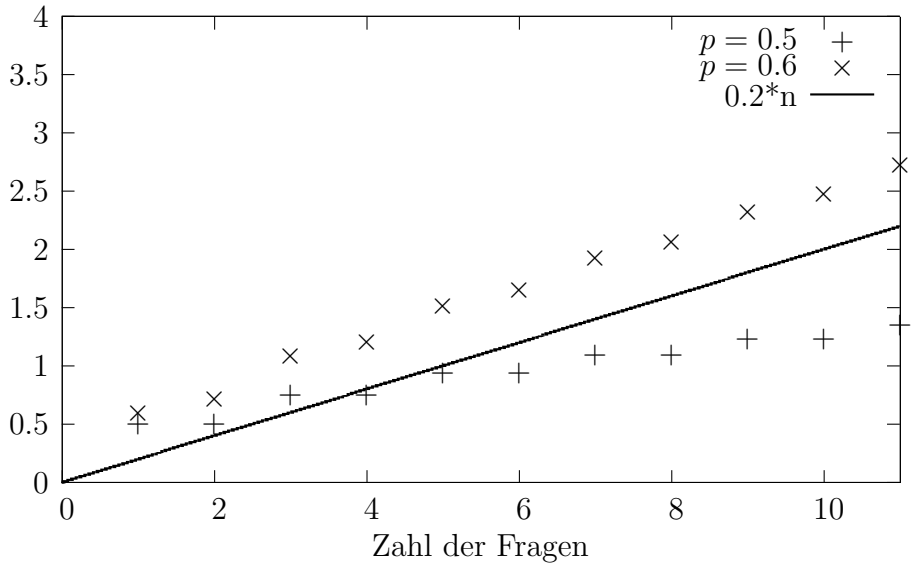

Abbildung 1: Erwartungswert für zwei Wahrscheinlichkeiten

ergibt sich exakt das selbe Ergebnis wie für n=10. Das gilt analog für alle ungeraden n, dass man das selbe Ergebnis auch für $n+1$ hat. Hat man nämlich n Fragen beantwortet, so bringt einem Raten in der n+1-sten Frage im Mittel nur dann weiter, wenn man gerade 0 Punkte hat (und zwar als Punktlandung, nicht genullte negativ-Punkte). Dann ändert sich nämlich bei Misserfolg nichts. Bei einer ungeraden Fragenzahl ist es aber nicht möglich, genau 0 Punkte zu haben.

Tatsächlich ist für $n = 11$:

$$\langle f \rangle = \left(1260 + \binom{10}{5} \cdot \frac{1}{2}\right) /1024 = (1260 + 126)/1024 = 2772/2048 \qquad (17)$$

Also wie bei $n = 10$ plus die Wahrscheinlichkeit der Nullpunkte-Punktlandung ($\binom{10}{5}/1024$) mal $\frac{1}{2}$ (der mittleren Punktezahl in einer Aufgabe). Der entsprechenden Summe (Gleichung 8 für $n = 11$) kann man das aber nicht ansehen, ich zumindest nicht.

Dieser Effekt verschwindet natürlich immer mehr, wenn man zu größeren Ratewahrscheinlichkeiten übergeht. Abbildung 1 zeigt die Erwartungswerte für $n < 12$ für zwei Wahrscheinlichkeiten.

Ab etwa $n = 30$ wird unser Programm extrem zäh und kurz danach berechnet es auch falsche Werte, weil der `int` für die Binomialkoeffizienten überläuft. Einige Optimierungen und Anpassungen ermöglichen aber eine weitere Berechnung, die in Abbildung 2 zu sehen ist.

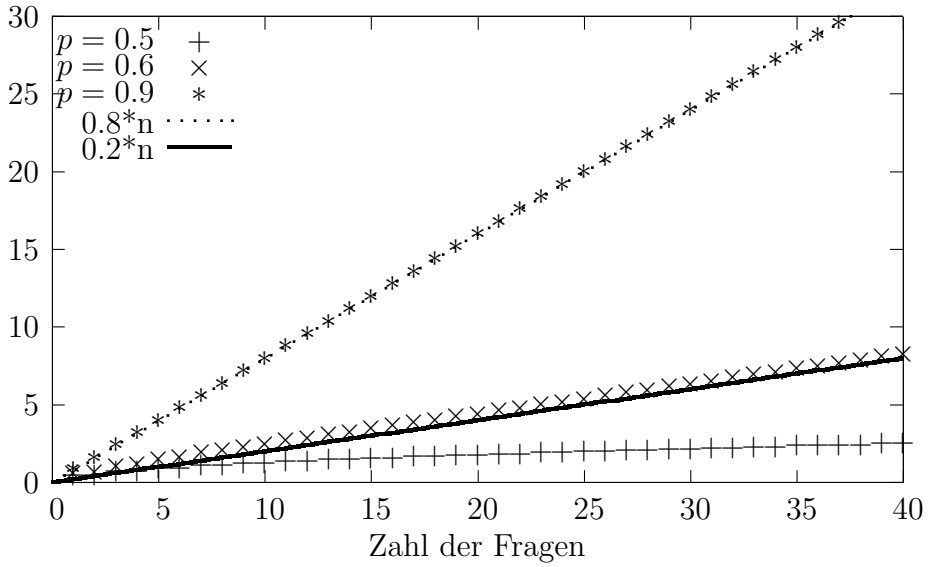

Abbildung 2: Erwartungswert für drei Wahrscheinlichkeiten

7 Asymptotik

Bei großen Wahrscheinlichkeiten p ist die Wahrscheinlichkeit negative Punkte zu bekommen (die dann auf Null gesetzt werden) klein und wir können die "Halb"–Summe 9 zur ganzen Summe ausdehnen und analytisch berechnen (wie in [1])

$$\langle f \rangle \approx \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (2k-n) = 2pn - n \tag{18}$$

Damit nähert sich die Kurve der Funktion $\langle f \rangle \approx p*n-(1-p)*n$. Im Übrigen sind die zusätzlichen Summanden alle negativ, es ist also stets $\langle f \rangle > p*n-(1-p)*n$.

Abbildung (1) und (2) zeigt diese Asymptotik zusammen mit den einzelnen Erwartungswerten. Abbildung (3) zeigt die Abweichung von der Asymptotik. Für $p = 0.9$ ist die Näherung sehr gut, aber auch für $p = 0.6$ wird die Näherung mit grösserem n immer besser.

Nur für $p = 0.5$ ist der dominante Effekt, dass negative Punkte auf Null gesetzt werden. Hier ist obige Abschätzung nicht besonders nützlich, und um 100% falsch, besagt sie doch nur $\langle f \rangle > 0$.

Für große n läßt sich die Binomialverteilung durch eine Normalverteilung annähern. Es gilt[1]:

$$B(k|p,n) \approx \frac{1}{\sqrt{2\pi npq}} \cdot \exp\left(-\frac{(k-np)^2}{2npq}\right) \tag{19}$$

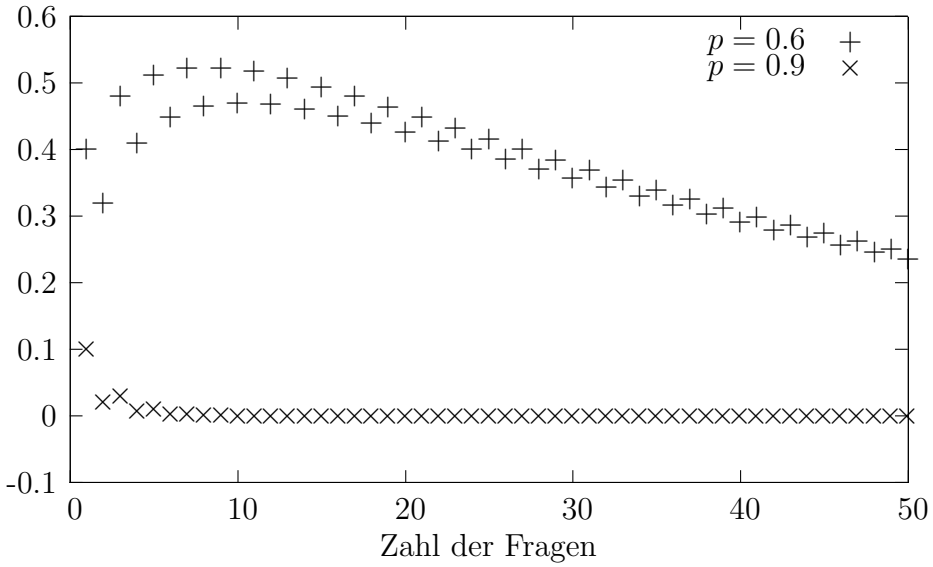

Abbildung 3: Abweichung des Erwartungswert von der Asymptotik

bzw. mit $p = q = \frac{1}{2}$

$$B(k|\frac{1}{2}, n) \approx \frac{\sqrt{2}}{\sqrt{\pi n}} \cdot \exp\left(-\frac{(2k-n)^2}{2n}\right). \tag{20}$$

und aus Gleichung (8) wird

$$\langle f \rangle = \frac{1}{2^n} \sum_{k=[\frac{n}{2}]+1}^{n} \binom{n}{k} \approx \int_{n/2}^{\infty} \frac{\sqrt{2}(2k-n)}{\sqrt{\pi n}} \exp\left(-\frac{(2k-n)^2}{2n}\right) dk \tag{21}$$

Mit der Substitution $u = -\frac{(2k-n)^2}{2n}$ läßt sich das Integral lösen und man erhält

$$\langle f \rangle \approx \sqrt{\frac{n}{2\pi}} \approx 0.399\sqrt{n}. \tag{22}$$

Abbildung (4) zeigt die gute Übereinstimmung mit der Asymptotik, die Erwartungswerte liegen abwechselnd darüber und darunter.

Der Übergang zur Normalverteilung funktioniert zwar auch für $p \neq 0.5$, aber das Integral läßt sich nicht mehr geschlossen lösen.

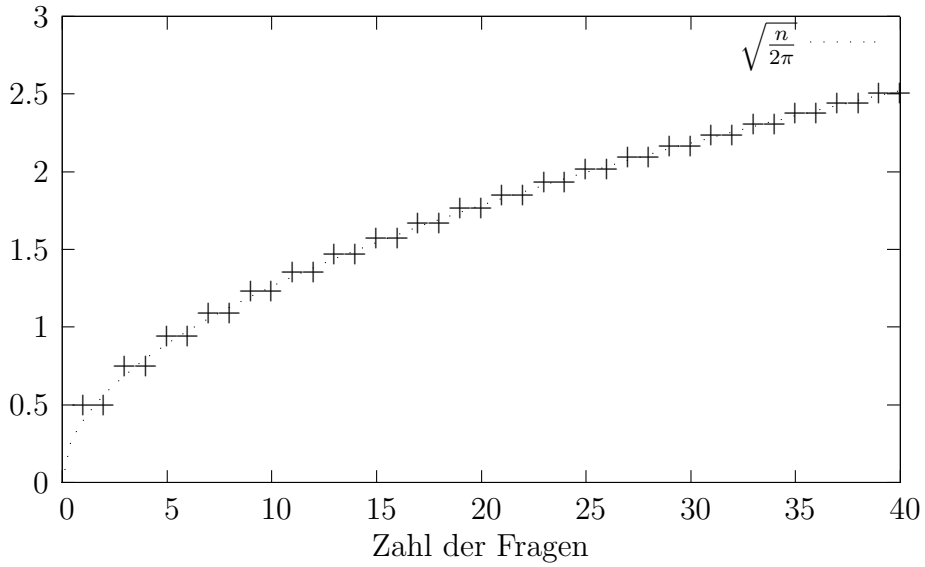

Abbildung 4: Asymptotik für $p = 0.5$

8 Summation und mittlere betragsmäßige Abweichung

Schließlich wollen wir uns noch einmal die Summe (8) vornehmen. Bei näherem Hinsehen findet sich eine Ähnlichkeit zur Berechnung der mittleren betragsmäßigen Abweichung oder mittleren absoluten Abweichung (englisch: mean absolute deviation[4]):

$$\mathrm{MAD} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} |k - np| \tag{23}$$

Für $p = 0.5$ ist nämlich $|k - np|$ gerade die Hälfte unserer Punktefunktion. Dafür wird die Summe, durch den Betrag, einmal über die untere und einmal über die obere Hälfte ausgeführt. Damit ist $\mathrm{MAD} = \langle f \rangle$. Darüber hinaus läßt sich mit Hilfe der Todthunterformel[2, 6, 7] die Summe $\sum \binom{n}{k} p^k (1-p)^{n-k} (k - np)$ für beliebige obere und untere Grenzen lösen. Schließlich stellt man auch noch fest, dass auch das Problem des 1–dimensionalen Random Walk[5] unserem Problem ähnelt. Die meisten dieser Ideen gehen allerdings auf de Moivre[3] zurück. Wer das Original lesen will, sei allerdings gewarnt: de Moivre veröffentlichte auf Latein und die Mathematik schrieb man auch sehr ungewohnt.

Von dieser Recherche lassen wir uns inspirieren und berechnen:

$$(k+1)\binom{n}{k+1} = (k+1)\frac{n!}{(k+1)!(n-k-1)!} = \frac{n!(n-k)}{k!(n-k)!} = (n-k)\binom{n}{k} \tag{24}$$

Dabei wurde der Binomialkoeffizient mit Hilfe der Fakultäten ausgedrückt, mit $k+1$ gekürzt, mit $n-k$ erweitert und schliesslich wieder der Binomialkoeffizient eingeführt.

Wegen $2k - n = k - (n - k)$ gilt

$$\begin{aligned}
\sum_{k=[\frac{n}{2}]+1}^{n} \binom{n}{k}(2k-n) &= \sum_{k=[\frac{n}{2}]+1}^{n} \binom{n}{k}k - \sum_{k=[\frac{n}{2}]+1}^{n-1} \binom{n}{k}(n-k) && (25)\\
&= \sum_{k=[\frac{n}{2}]+1}^{n} \binom{n}{k}k - \sum_{k=[\frac{n}{2}]+1}^{n-1} \binom{n}{k+1}(k+1) && (26)\\
&= \binom{n}{[\frac{n}{2}]+1}([\frac{n}{2}]+1) && (27)
\end{aligned}$$

Offenbar heben sich die beiden Summen gegenseitig weg, bis auf den ersten Term der ersten Summe. Den letzten Term der zweiten Summe hatten wir schon im ersten Schritt weggelassen, da er Null ist. Im zweiten Schritt hatten wir natürlich die Beziehung (24) verwendet.

Aus unserem Erwartungswert (Gleichung (8)) wird daher:

$$\langle f\rangle = \frac{1}{2^n}\binom{n}{[\frac{n}{2}]+1}([\frac{n}{2}]+1) \tag{28}$$

Wenn wir die Fälle für gerade und ungerade n unterscheiden, können wir noch weiter vereinfachen. Wir betrachten zunächst gerade n also $n = 2j$:

$$\langle f\rangle = \frac{1}{2^n}\binom{n}{\frac{n}{2}}\frac{n}{2} = \frac{1}{2^{2j}}\frac{(2j)!}{j!j!}j = \left(\frac{1}{2^j}\right)^2\frac{(2j)!}{j!(j-1)!} \tag{29}$$

Durch Einführung der Doppelfakultät $(2j-2)!! = 2^{j-1}(j-1)!$ und $(2j-1)!! = \frac{(2j)!}{2^j j!}$ erhält man

$$\langle f\rangle = \frac{1}{2}\frac{(2j-1)!!}{(2j-2)!!} = \frac{1}{2}\frac{(n-1)!!}{(n-2)!!} \tag{30}$$

Setzt man in Gleichung (29) die Stirlingformel $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$ ein, erhalten wir wieder die Asymptotik (22).

Für ungerade $n = 2j + 1$ erhält man analog

$$\langle f\rangle = \frac{1}{2^n}\binom{n}{\frac{n+1}{2}}\frac{n+1}{2} = \frac{(2j+1)!(j+1)}{2^{2j+1}(j+1)!j!} = \frac{1}{2}\frac{(2j+1)!!}{(2j)!!} = \frac{1}{2}\frac{n!!}{(n-1)!!} \tag{31}$$

Hier kann man leicht erkennen, dass der Wert für das nachfolgende gerade n genau gleich groß ist.

9 Zusammenfassung

Wir haben hier ein nicht so ganz triviales Problem der Kombinatorik behandelt. Die Binomialkoeffizienten bzw. die Binomialverteilung ist hier das entscheidende Element.

Für hinreichend große Ratewahrscheinlichkeiten ist alles einfach, die Zahl der Punkte hängt linear von der Ratewahrscheinlichkeit ab. Nur für reines Raten ist die Asymmetrie, die aus der Unmöglichkeit negativer Punkte resultiert entscheidend, aber auch für große Ratewahrscheinlichkeiten ist Raten im Schnitt besser als das Auslassen einer Aufgabe.

Die Frage nach dem Erwartungswert beim reinen Raten ist beantwortet. Ein Schnitt von 1.23 Punkten ist bei 10 Fragen zu erwarten. Tatsächlich lag der Schnitt bei etwa 2.632 Punkten, also haben doch nicht nur alle geraten. Wenn jeder alle Fragen beantwortet hätte, entspräche dies einer Ratewahrscheinlichkeit von etwa 61%. Insgesamt sollte man den Studierenden raten, nur dann eine Frage auszulassen, wenn sie sich ganz sicher sind, die anderen Fragen richtig beantwortet zu haben, was für die Mehrheit offensichtlich nicht zutrifft. Raten hilft also weiter, wenn man nichts weiß, und schadet auch nicht (im Mittel!), wenn man etwas weiß.

Allerdings gibt es in der Praxis keine einheitliche Ratewahrscheinlichkeit. Bei manchen Fragen ist man sich sicher, bei anderen weniger und bei manchen weiss man gar nichts.

Für große Fragenzahl n läßt sich die Asymptotik

$$\langle f\rangle = \sum_{k=[\frac{n}{2}]+1}^{n} \binom{n}{k} p^k (1-p)^{n-k} (2k-n) \underset{n\to\infty}{\approx} \begin{cases} 2pn-n & p > \frac{1}{2} \\ \sqrt{n/2\pi} & p = \frac{1}{2} \end{cases} \tag{32}$$

angeben. Schliesslich ist für $p = 0.5$ durch Vergleich mit der Berechnung der mittleren absoluten Abweichung, oder dem eindimensionalen Random Walk, eine geschlossene Berechnung der Summe möglich.

Interessant wäre diese Ideen auf $p > 0.5$ auszudehnen oder wenigstens die Abweichung von der Asymptotik zu schätzen. Ich vermute ja sowas wie $\langle f\rangle \approx 2pn - n + \alpha\sqrt{np^{\beta n}}$ mit geeigneten Parametern α und β. Vielleicht sollte man auch mal darüber nachdenken was bei $p < 0.5$ passiert, obwohl das absichtlich falsch Raten für die ursprüngliche Aufgabenstellung sinnlos ist.

A C Code zur Berechnung $p = 0.5$

```
/* cc -o erwartungswert erwartungswert.c -lm */
#include<stdio.h>
#include<math.h>

int binomial(int n,int k){
   if( k>=n) return 1;
   if(k==0) return 1;
   return binomial(n-1,k)+binomial(n-1,k-1);
}

int Halbsumme(int n){
   int i,hs=0,punkte;
   for(i=0;i<=n;i++){
        if(i>n/2) punkte = 2*i-n; else punkte=0;
        printf("   %d richtige: %d mal %d Punkte = %d \n",
                   i,   binomial(n,i), punkte, binomial(n,i)*punkte);
        hs+=binomial(n,i)*punkte;
   }
   return hs;
}

int main(){
   int n,hs,gesamtzahl;
   double erwart;
   for(n=1;n<=10;n++){
      printf("%d Fragen:\n",n);
      hs=Halbsumme(n);
      gesamtzahl=(int) pow(2,n);
      erwart= ((double) hs)/pow(2,n);
      printf("Erwartungswert bei %d Fragen %d/%d = %f\n\n",
                                  n,    hs,gesamtzahl,erwart);
   }
}
```

Literatur

[1] http://de.wikipedia.org/wiki/Binomialverteilung

[2] Persi Diaconis and Sandy Zabell *Closed Form Summation for Classical Distributions: Variations on a Theme of De Moivre*, Statistical Science, **6**, No. 3 (Aug., 1991), pp. 284-302

[3] de Moivre A., *Miscellanea Analytica* (1730)

[4] Weisstein, Eric W. "Mean Deviation." From MathWorld, a Wolfram Web Resource. http://mathworld.wolfram.com/MeanDeviation.html

[5] Weisstein, Eric W. "Random Walk 1-Dimensional" From MathWorld, a Wolfram Web Resource.
http://mathworld.wolfram.com/RandomWalk1-Dimensional.html

[6] S. L. Zabell *Symmetry and Its Discontents: Essays on the History of Inductive Probability*

[7] Todthunter I., *A History of the Mathematical Theory of Probability*, Macmillam London (1865)