Surveys in Combinatorics, 1989 (ed. J. Siemons)

The 'Snake Oil' method for proving combinatorial identities

Herbert S. Wilf*

I have recently been working on a book about generating functions. It will be called 'Generatingfunctionology,' and it is intended to be an upper-level undergraduate, or graduate text in the subject. The object is to try to impress students with the beauty of this subject too, so they won't think that only bijections can be lovely.

In one section of the book I will discuss combinatorial identities, and the approach will be this. First I'll give the 'Snake Oil' method, in the spirit of a unified approach that works on many relatively simple identities. It involves generating functions. Second, I will write about a much more powerful method that works on 'nearly all' identities, including all classical hypergeometric identities and many, many binomial coefficient identities.

These two approaches will be mirrored here, in that this article will mostly be about the Snake Oil method, whereas the talk that I will give at the conference will be about the much more powerful method of WZ pairs [WZ], which at this writing is still under development. A brief summary of the WZ results appears in section (II) below.

Aside from these developments, there have been other unifying forces at work in the field of identities. The expository paper of Roy [Ro], shows how even without a computer one can recognize many binomial identities as cases of just a very few identities in the theory of hypergeometric series. The work of Knuth [Kn] shows how a few rules about binomial coefficients and their handling can, in skilled hands, prove many difficult identities. This point is reinforced in [GKP].

A third method of great generality is due to Egorychev [Eg], and is called by him the 'method of residues.' It is quite strong on binomial coefficient identities, but does not have the scope of the WZ machine.

(I) The 'Snake Oil' method

The first method that I want to discuss here is the 'Snake Oil' method, after a slang expression for a quack medicine or nostrum,† that is highly touted as a panacæa, but in fact falls quite short. I hope you won't think it falls quite short, because I'm not touting it as a cure-all. It is, though, an excellent medicine to try

* Research supported in part by the United States Office of Naval Research

† The Random House Dictionary of the English Language states that 'snake oil' is of American origin. ca. 1925-30, and that a typical use of the expression is *The governor promised to lower taxes, but it was the same old snake oil.*

for whatever sum ails you.

The ingredients of the method are the following.

(a) One needs to observe the convention that the binomial coefficient $\binom{n}{k}$ vanishes if $k < 0$, or if n and k are both integers and $0 \le n < k$.

(b) One needs to observe the convention that if a certain summation extends over an index k, say, and if the range of summation is not specified, then that range is from $-\infty$ to ∞.

(c) One needs to believe the binomial theorem

$$\sum_k \binom{n}{k} x^k = (1+x)^n, \tag{1}$$

which is valid for all x if n is a nonnegative integer, and for $|x| < 1$ otherwise.

(d) One needs the series expansion

$$\sum_{k \ge 0} \binom{k}{b} x^k = \frac{x^b}{(1-x)^{b+1}} \tag{2}$$

which is valid for nonnegative integer b and $|x| < 1$. [use $\binom{-r}{k} = (-)^k \binom{r+k-1}{k}$.]

The basic idea of the method is what I might call the *external* approach to identities rather than the usual *internal* method.

To explain the difference between these two points of view, suppose we want to prove some identity that involves binomial coefficients. Typically such a thing would assert that some fairly intimidating-looking sum is in fact equal to such-and-such a simple function of n (an excellent collection of these is Gould [Gou]).

One approach that is now customary consists primarily in looking inside the summation sign ('internally'), and using binomial coefficient identities or other manipulations of indices *inside* the summations, to bring the sum to manageable form.

In the *external*, or generatingfunctionological, approach, one begins by giving just a quick glance at the expression that is inside the summation sign, long enough to spot the 'free variables,' i.e., what it is that the sum depends on after the dummy variables have been summed over. Suppose that such a free variable is called n.

Then instead of trying to grapple with the sum, the trick is to find the generating function whose coefficients are the values of your sum. More precisely, if $f(n)$ is your sum, instead of going *inside* to evaluate $f(n)$, go *outside* to evaluate $\sum_n f(n)x^n$. Then after you've done that, read off the coefficient of x^n, and you're finished. Here it is, a little more systematically:

(a) Identify the free variable, say n, that the sum depends on. Give a name to the sum that you are working on: call it $f(n)$.

(b) Let $F(x)$ be the ordinary power series generating function (opsgf) whose coefficient of x^n is $f(n)$, the sum that you'd like to evaluate.

(c) Multiply the sum by x^n, and sum on n. Your generating function is now expressed as a double sum, over n, and over whatever variable was first used as a dummy summation variable.

(d) Interchange the order of the two summations that you are now looking at, and perform the inner one in simple closed form. For this purpose it will be helpful to have a catalogue of series whose sums are known.

(e) Try to identify the coefficents of the generating function of the answer, because those coefficients are what you want to find.

Several examples

1. A Fibonacci sum

Consider the sum

$$\sum_{k\geq 0}\binom{k}{n-k} \qquad (n=0,1,2,\ldots).$$

The free variable is n, so let's call the sum $f(n)$. Write it out like this:

$$f(n)=\sum_{k\geq 0}\binom{k}{n-k}.$$

Multiply both sides by x^n and sum over $n\geq 0$. You have now arrived at step (c) of the general method, and you are looking at

$$F(x)=\sum_{n\geq 0}x^n\sum_{k\geq 0}\binom{k}{n-k}.$$

Ready for step (d)? Interchange the sums, to get

$$F(x)=\sum_{k\geq 0}\sum_{n\geq 0}\binom{k}{n-k}x^n.$$

We would like to 'do' the inner sum, the one over n. The trick is to get the exponent of x to be exactly the same as the index that appears in the binomial coefficient. In this example the exponent of x is n, and n is involved in the downstairs part of the binomial coefficient in the form $n-k$. To make those the same, the correct medicine is to multiply inside the sum by x^{-k} and outside the inner sum by x^k, to compensate. The result is

$$F(x)=\sum_{k\geq 0}x^k\sum_{n\geq 0}\binom{k}{n-k}x^{n-k}.$$

Now the exponent of x is the same as the lower index of the binomial coefficient. Hence take $r=n-k$ as the new dummy variable of summation in the inner sum. We find then

$$F(x)=\sum_{k\geq 0}x^k\sum_{r}\binom{k}{r}x^r.$$

We recognize the inner sum immediately, as $(1+x)^k$. Hence

$$F(x)=\sum_{k\geq 0}x^k(1+x)^k=\sum_{k\geq 0}(x+x^2)^k=\frac{1}{1-x-x^2}.$$

The generating function on the right is an old friend; it generates the Fibonacci numbers (if you didn't recognize that, a partial fraction expansion would produce the coefficients quite explicitly). Hence $f(n)=F_n$, and we have discovered that

$$\sum_{k\geq 0}\binom{k}{n-k}=F_n \qquad (n=0,1,2,\ldots).$$

2. A bit harder

Consider the sum

$$\sum_{k\geq 0}\binom{n+k}{m+2k}\binom{2k}{k}\frac{(-1)^k}{k+1} \qquad (m,n\geq 0). \tag{3}$$

Let $f(n)$ denote the sum in question (m would work just as well), and let $F(x)$ be its opsgf. Dive in immediately by multiplying by x^n and summing over $n\geq 0$, to get

$$\begin{aligned}
F(x)&=\sum_{n\geq 0}x^n\sum_{k\geq 0}\binom{n+k}{m+2k}\binom{2k}{k}\frac{(-1)^k}{k+1}\\
&=\sum_{k\geq 0}\binom{2k}{k}\frac{(-1)^k}{k+1}x^{-k}\sum_{n\geq 0}\binom{n+k}{m+2k}x^{n+k}\\
&=\sum_{k\geq 0}\binom{2k}{k}\frac{(-1)^k}{k+1}x^{-k}\frac{x^{m+2k}}{(1-x)^{m+2k+1}} \qquad \text{(by (2))}\\
&=\frac{x^m}{(1-x)^{m+1}}\sum_{k\geq 0}\binom{2k}{k}\frac{1}{k+1}\left\{\frac{-x}{(1-x)^2}\right\}^k\\
&=\frac{-x^{m-1}}{2(1-x)^{m-1}}\left\{1-\sqrt{1+\frac{4x}{(1-x)^2}}\right\}\\
&=\frac{-x^{m-1}}{2(1-x)^{m-1}}\left\{1-\frac{1+x}{1-x}\right\}\\
&=\frac{x^m}{(1-x)^m}.
\end{aligned}$$

The original sum is therefore the coefficient of x^n in the last member above. But that is $\binom{n-1}{m-1}$, by (2) again, and we have our answer.

If the manipulations seemed long, consider that at least they're always the *same* manipulations, whenever the method is used, and also that with some effort a computer could be trained to do them!

3. An equality

Suppose we have two complicated sums and we want to show that they're the same. Then the generating function method, if it works, should be very easy to carry out. Indeed, one might just find the generating functions of each of the two sums independently and observe that they are the same.

Suppose we want to prove that

$$\sum_k \binom{m}{k}\binom{n+k}{m} = \sum_k \binom{m}{k}\binom{n}{k}2^k \qquad (n,m \geq 0),$$

without evaluating either of them.

Multiply on the left by x^n, sum on $n \geq 0$ and interchange the summations, to arrive at

$$\begin{aligned}\sum_k \binom{m}{k}x^{-k}\sum_{n\geq 0}\binom{n+k}{m}x^{n+k} &= \sum_k \binom{m}{k}x^{-k}\frac{x^m}{(1-x)^{m+1}} \\ &= \frac{x^m}{(1-x)^{m+1}}\left(1+\frac{1}{x}\right)^m \\ &= \frac{(1+x)^m}{(1-x)^{m+1}}.\end{aligned}$$

If we multiply on the right by x^n, etc., we find

$$\begin{aligned}\sum_k \binom{m}{k}2^k\sum_{n\geq 0}\binom{n}{k}x^n &= \frac{1}{(1-x)}\sum_k \binom{m}{k}\left(\frac{2x}{(1-x)}\right)^k \\ &= \frac{1}{(1-x)}\left(1+\frac{2x}{1-x}\right)^m \\ &= \frac{(1+x)^m}{(1-x)^{m+1}}.\end{aligned}$$

Since the generating functions are the same, the sums are equal, and we're finished.

4. An identity of D. E. Knuth

Here we want to evaluate, for $m, n \geq 0$,

$$\sum_k \binom{n}{k}\binom{n-k}{\lfloor\frac{m-k}{2}\rfloor}y^k. \tag{4}$$

The appearance of the 'floor' function is a bit disquieting, and in fact, it removes this sum from the realm of identities of hypergeometric type. Nevertheless, Snake Oil is more than equal to the occasion.

We are going to multiply the sum by x-to-the-something, and sum over the 'something'. There are two choices: m, or n, because there are two free variables. The one to use is m, because it appears in only one place inside the sum, hence after

the step of interchanging orders of summation, that sum will have only one binomial coefficient in it.

Call the unknown sum $f_m(y)$. Multiply (4) by x^m and sum on $m \geq 0$, to get

$$F(x;y) = \sum_k \binom{n}{k}y^k \sum_{m\geq 0}\binom{n-k}{\lfloor\frac{m-k}{2}\rfloor}x^m.$$

The next step is that inside the sum we make the exponent of x into $m-k$, to obtain

$$\begin{aligned}F(x;y) &= \sum_k \binom{n}{k}(xy)^k \sum_{m\geq 0}\binom{n-k}{\lfloor\frac{m-k}{2}\rfloor}x^{m-k} \\ &= \sum_k \binom{n}{k}(xy)^k \sum_{r\geq 0}\binom{n-k}{\lfloor r/2\rfloor}x^r.\end{aligned}$$

This problem differs from the preceding ones in that we don't know offhand about the inner sum. But it's not hard to find that

$$\begin{aligned}\sum_{r\geq 0}\binom{q}{\lfloor r/2\rfloor}x^r &= \binom{q}{0} + \binom{q}{0}x + \binom{q}{1}x^2 + \binom{q}{1}x^3 + \cdots \\ &= (1+x)(1+x^2)^q \qquad (q \geq 0).\end{aligned}$$

Putting it all together, our unknown sum (4) is the coefficient of x^m in

$$\begin{aligned}F(x;y) &= (1+x)\sum_k \binom{n}{k}(xy)^k(1+x^2)^{n-k} \\ &= (1+x)(1+xy+x^2)^n.\end{aligned}$$

That's as far as we can go without specifying y. If $y = \pm 2$ we can go further, to get

$$\sum_k \binom{n}{k}\binom{n-k}{\lfloor\frac{m-k}{2}\rfloor}2^k = \binom{2n+1}{m}$$

and

$$\sum_k \binom{n}{k}\binom{n-k}{\lfloor\frac{m-k}{2}\rfloor}(-2)^k = (-1)^m\binom{2n}{m}\left\{\frac{2n-2m+1}{2n-m+1}\right\}.$$

5. Not necessarily binomial coefficients

The Snake Oil method works not only on sums that involve binomial coefficients, but on all sorts of counting numbers, as this example shows.

Let $\{a_n\}$ and $\{b_n\}$ be two sequences whose exponential generating functions are respectively $A(x)$, $B(x)$. Suppose that the sequences are connected by

$$b_n = \sum_k s(n,k)a_k \qquad (n \geq 1), \tag{5}$$

where the s's are the Stirling numbers of the first kind. Let $A(x)$ and $B(x)$ denote the exponential generating functions of these sequences, and let's find out how $A(x)$ and $B(x)$ are related to each other, in view of (5).

To do that, multiply (5) by $x^n/n!$, sum over n and use the fact that the Stirling numbers satisfy

$$\sum_n \frac{s(n,k)}{n!}x^n = \frac{1}{k!}\left\{\log\frac{1}{1-x}\right\}^k \qquad (k \geq 0).$$

The result is that

$$\begin{aligned} B(x) &= \sum_n \frac{x^n}{n!}\sum_k s(n,k)a_k \\ &= \sum_{k\geq 0} a_k \left\{\log\frac{1}{1-x}\right\}^k \\ &= A\left(\log\frac{1}{1-x}\right). \end{aligned} \tag{6}$$

Hence the exponential generating function relation (6) is equivalent to the sequence relation (5). As an example of its use, take the $\{a_n\}$ to be the Bernoulli numbers $\{B_n\}$, so that $A(x) = x/(e^x - 1)$. Then (6) says that

$$B(x) = \frac{1-x}{x}\log\frac{1}{1-x} = -\sum_{n\geq 1}\frac{x^n}{n(n+1)},$$

which is clearly the exponential generating function of the sequence

$$\left\{-\frac{n!}{n(n+1)}\right\}_{n\geq 1}.$$

Consequently we have the identity

$$\sum_k s(n,k)B_k = -\frac{n!}{n(n+1)} \qquad (n \geq 1)$$

between the Bernoulli numbers and the Stirling numbers of the first kind.

6. A tough one

For our final example, we prove a difficult identity of Graham and Riordan, namely

$$f(m) = \sum_k \binom{2n+1}{2k}\binom{m+k}{2n} = \binom{2m+1}{2n} \qquad (m,n \geq 0). \tag{7}$$

First, notice that we singled out the variable m as the free variable, rather than n, which is also a free variable. We did that because the m appears in only one place on the left side of (7), while n appears in two places. That means that after we multiply by x^m, sum, and interchange the summations, the sum on m will have just a single binomial coefficient in it, rather than two.

This is a general characteristic of the Snake Oil method. It works best if there is a free variable that appears only once.

Next, we would normally multiply through in (7) by x^m and continue, but since the claimed answer has $2m+1$ in it, it might save time if we multiply by x^{2m+1} instead (it only saves time; the method works fine if we use x^m, but it's a little more complicated to use).

If we do that, the result is that the ordinary power series generating function $F(x)$ of the sums on the left side of (7) is

$$\begin{aligned} \sum_{m\geq 0} f(m)x^{2m+1} &= \sum_{m\geq 0} x^{2m+1}\sum_k \binom{2n+1}{2k}\binom{m+k}{2n} \\ &= \sum_k \binom{2n+1}{2k}\sum_{m\geq 0}\binom{m+k}{2n}x^{2m+1} \\ &= \sum_k \binom{2n+1}{2k}x^{-2k+1}\sum_{m\geq 0}\binom{m+k}{2n}x^{2(m+k)} \\ &= \sum_k \binom{2n+1}{2k}x^{-2k+1}\sum_r \binom{r}{2n}x^{2r}. \end{aligned}$$

The innermost sum is, by (2),

$$\frac{x^{4n}}{(1-x^2)^{2n+1}},$$

and therefore

$$\begin{aligned} F(x) &= \frac{x^{4n+1}}{(1-x^2)^{2n+1}}\sum_k \binom{2n+1}{2k}x^{-2k} \\ &= \frac{x^{4n+1}}{(1-x^2)^{2n+1}}\left\{\frac{1}{2}(1+x^{-1})^{2n+1} + \frac{1}{2}(1-x^{-1})^{2n+1}\right\} \\ &= \frac{x^{2n}}{2}\left\{\frac{1}{(1-x)^{2n+1}} - \frac{1}{(1+x)^{2n+1}}\right\}. \end{aligned}$$

We are now finished, since the coefficient of x^{2m+1} in the last member is clearly as shown on the right side of (7).

In general, whenever a free variable appears only once inside an unknown sum, Snake Oil may be the best medicine. The list of identities that can be handled by this simple, programmable device is very lengthy indeed.

(II) WZ Pairs

Here is a brief summary of the method of WZ pairs, which at the time of this writing is just being brought into final form.

The method provides machinery whereby a vast collection of combinatorial identities can be given one-line proofs. The proof certificate consists of a single rational function. Here is an example of a one-line proof of a famous and difficult identity.

Theorem (Dixon). *The identity*

$$\sum_k (-1)^k \binom{n+b}{n+k}\binom{n+c}{c+k}\binom{b+c}{b+k} = \frac{(n+b+c)!}{n!b!c!}$$

is true.

Proof: Take

$$R(n,k) = \frac{(c+1-k)(b+1-k)}{2(n+k)(n+b+c+1)}. \blacksquare$$

Here is a one-line proof of the famous hypergeometric identity of Saalschütz.

Theorem. *The identity*

$$\sum_k \frac{(a+k-1)!(b+k-1)!(n-k-a-b+c-1)!}{k!(n-k)!(k+c-1)!} = \frac{(n+c-a-1)!(n+c-b-1)!}{n!(n+c-1)!}.$$

is true.

Proof: Take

$$R(n,k) = -\frac{(b+k-1)(a+k-1)}{(c-b+n)(c-a+n)}. \blacksquare$$

Now here is how to complete the proof of an identity if such a rational function be given. First, divide through the identity by its right hand side, so it takes the standard form $\sum_k F(n,k) = 1 \quad (n \geq 0)$. Next, define a function $G(n,k) = R(n,k)F(n,k-1)$, where R is the rational function that is given in the 'one-line' proof certificate. Then verify that the pair (F,G) satisfy the WZ condition

$$F(n+1,k) - F(n,k) = G(n,k+1) - G(n,k).$$

Finally check that the boundary conditions $G(n,\pm\infty) = 0$ are satisfied.

Having done that, you have proved that $\sum_k F(n,k) = const.$, for all $n \geq 0$, i.e., you have proved the identity.

Since the process of verifying the proof certificate is so mechanical, the whole job can be done by computer, and the task of proving complicated identities can be removed from the list of human tasks.

Such rational function certificates exist for all of the classical hypergeometric identities, such as those of Saalschütz, Dixon, Whipple, Dougall, Clausen, and for a host of newer hypergeometric identities as well. As is well known, these hypergeometric identities imply huge numbers of binomial coefficient identities, and so all of these have certificates too. For more details see [WZ].

References

[Eg] Egorychev, G.P., Integral representation and the computation of combinatorial sums, American Mathematical Society Translations, vol. 59, 1984.

[GKP] Graham, Ronald L., Knuth, Donald. E., and Patashnik, Oren, Concrete Mathematics, Addison-Wesley, 1989.

[Gou] Gould, Henry W., Combinatorial Identities, Morgantown, W. Va., 1972.

[Go] Gosper, R. William, Jr., Decision procedure for indefinite hypergeometric summation, Proc. Nat. Acad. Sci. U. S. A. **75** (1978), 40-42.

[Kn] Knuth, Donald E., The Art of Computer Programming, vol. 1: Fundamental Algorithms, 1968 (2nd ed. 1973); vol. 2: Seminumerical Algorithms, 1969 (2nd ed. 1981); vol. 3: Sorting and Searching, 1973, Addison-Wesley.

[Ro] Roy, Ranjan, Binomial identities and hypergeometric series, Amer. Math. Monthly **97** (1987), 36-46.

[WZ] Wilf, Herbert S., and Zeilberger, Doron, WZ pairs certify combinatorial identities, preprint.

Department of Mathematics
University of Pennsylvania
Philadelphia, PA 19104-6395