

The Curious Incident of the Combinatorial Identity

Sudeep Kamath
sudeep@eecs.berkeley.edu
www.eecs.berkeley.edu/~sudeep

October 14, 2010

Abstract

Under an unoriginal title, this short write-up documents an informal account of how I stumbled upon an interesting combinatorial identity.

A Fun Problem

During the summer of 2007, I was a lowly undergraduate intern at Stanford University working with Prof. Balaji Prabhakar. Our meetings used to be a lot of fun and in one of them, he suggested to me the following problem.

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. random variables, distributed exponentially with mean 1. Define $S_m := \frac{1}{m} \sum_{i=1}^m X_i$ and $Z_n := \max_{1 \leq m \leq n} S_m$. Show that

$$\mathbb{E}[Z_n] = \sum_{k=1}^n \frac{1}{k^2} .$$

Not yet fully jaded at the time, I found myself with the naïve eagerness that undergraduates may be familiar with.

Brute Force

Let $F_{Z_n}(t) := \Pr(Z_n \leq t)$. Then,

$$\begin{aligned} F_{Z_{n+1}}(t) &= \Pr(S_1 \leq t, S_2 \leq t, \dots, S_{n+1} \leq t) \\ &= \Pr(X_1 \leq t, X_2 \leq 2t - X_1, \dots, X_{n+1} \leq (n+1)t - X_1 - X_2 - \dots - X_n) \\ &= \int_0^t \int_0^{2t-x_1} \dots \int_0^{(n+1)t - \sum_{j=1}^n x_j} e^{-(x_1+x_2+\dots+x_n+x_{n+1})} dx_{n+1} \dots dx_2 dx_1 \\ &= \int_0^t \int_0^{2t-x_1} \dots \int_0^{nt - \sum_{j=1}^{n-1} x_j} e^{-(x_1+x_2+\dots+x_n)} \left[1 - e^{-((n+1)t - \sum_{j=1}^n x_j)}\right] dx_n \dots dx_2 dx_1 \\ &= F_{Z_n}(t) - e^{-(n+1)t} V_n(t) , \end{aligned}$$

where $V_n(t)$ is the following volume integral:

$$V_n(t) := \int_0^t \int_0^{2t-x_1} \int_0^{3t-x_1-x_2} \dots \int_0^{nt-x_1-x_2-\dots-x_{n-1}} dx_n dx_{n-1} \dots dx_2 dx_1 .$$

By a dimension argument, we have $V_n(t) = c_n t^n$. Now,

$$\begin{aligned}\mathbb{E}[Z_{n+1}] &= \int_0^\infty [1 - F_{Z_{n+1}}(t)] dt \\ &= \int_0^\infty [1 - F_{Z_n}(t) + e^{-(n+1)t} c_n t^n] dt \\ &= \mathbb{E}[Z_n] + \frac{c_n n!}{(n+1)^{n+1}}\end{aligned}$$

As $\mathbb{E}[Z_1] = \mathbb{E}[X_1] = 1$, we only need to show that $\frac{c_n n!}{(n+1)^{n+1}} = \frac{1}{(n+1)^2}$, or that

$$c_n = \frac{(n+1)^{n-1}}{n!}.$$

Generalize!

After a few unsuccessful hacks on to it, I recognized that sometimes it is easier to solve a more general problem. I was thus, led to defining

$$V_{n,r}(t) := \int_0^t \int_0^{2t-x_1} \int_0^{3t-x_1-x_2} \dots \int_0^{nt-\sum_{j=1}^{n-1} x_j} (x_1 + x_2 + \dots + x_n)^r dx_n dx_{n-1} \dots dx_2 dx_1,$$

where $V_n(t) = V_{n,0}(t)$. Again, by a dimension argument, $V_{n,r}(t) = c_{n,r} t^{n+r}$ with $c_{n,0} = c_n$. By the transformation $y_i = \sum_{j=1}^i x_j$ for $1 \leq i \leq n$ (which has Jacobian 1), we have

$$V_{n,r}(t) = \int_0^t \int_{y_1}^{2t} \int_{y_2}^{3t} \dots \int_{y_{n-1}}^{nt} (y_n)^r dy_n dy_{n-1} \dots dy_2 dy_1.$$

A Mistake Well-Made

I proceeded ruthlessly with the integrals. Note: (1) to (2) is actually wrong! I didn't notice it at the time though.

$$V_{n,r}(t) = \int_0^t \int_{y_1}^{2t} \int_{y_2}^{3t} \dots \int_{y_{n-1}}^{nt} (y_n)^r dy_n dy_{n-1} \dots dy_2 dy_1 \tag{1}$$

$$= \int_0^t \int_{y_1}^{2t} \int_{y_2}^{3t} \dots \int_{y_{n-2}}^{(n-1)t} \frac{(nt - y_{n-1})^{r+1}}{r+1} dy_{n-1} \dots dy_2 dy_1 \tag{2}$$

$$= \int_0^t \int_{y_1}^{2t} \int_{y_2}^{3t} \dots \int_{y_{n-3}}^{(n-2)t} \frac{(nt - y_{n-2})^{r+2}}{(r+1)(r+2)} dy_{n-2} \dots dy_2 dy_1 - \frac{t^{r+2}}{(r+1)(r+2)} V_{n-2}(t)$$

$$= \int_0^t \int_{y_1}^{2t} \int_{y_2}^{3t} \dots \int_{y_{n-4}}^{(n-3)t} \frac{(nt - y_{n-3})^{r+3}}{(r+1)(r+2)(r+3)} dy_{n-3} \dots dy_2 dy_1$$

$$- \frac{2^{r+3} t^{r+3} V_{n-3}(t)}{(r+1)(r+2)(r+3)} - \frac{t^{r+2} V_{n-2}(t)}{(r+1)(r+2)}$$

$$= \int_0^t \frac{(nt - y_1)^{n+r-1}}{(r+1)(r+2) \dots (r+n-1)} dy_1 - \sum_{k=2}^{n-1} \frac{(k-1)^{k+r} t^{k+r} V_{n-k}(t)}{(r+1)(r+2) \dots (r+k)}$$

$$= \frac{r! (n^{n+r} - (n-1)^{n+r}) t^{n+r}}{(n+r)!} - \sum_{k=2}^{n-1} \frac{r! (k-1)^{k+r} t^{k+r} V_{n-k}(t)}{(k+r)!}$$

This gives us an (incorrect) relation among the $c_{n,r}$'s:

$$c_{n,r} = \frac{r!(n^{n+r} - (n-1)^{n+r})}{(n+r)!} - \sum_{k=2}^{n-1} \frac{r!(k-1)^{k+r} c_{n-k}}{(k+r)!} . \quad (3)$$

For $r = 0$, this gives (using $c_0 = 1$ by convention, consistent with the formula)

$$\sum_{k=0}^n \frac{(k-1)^k}{k!} \cdot c_{n-k} = \frac{n^n}{n!} . \quad (4)$$

Plugging in $c_l = \frac{(l+1)^{l-1}}{l!}$ for each l and verifying the equality thus obtained would complete the proof by induction. Given some complicated n -fold integral, one cannot be too sure that one would have much to say about it. But a combinatorial identity? How hard can that be? Turns out that we have got to verify this:

$$\sum_{k=0}^n \binom{n}{k} (k-1)^k (n-k+1)^{n-k-1} = n^n . \quad (5)$$

MATLAB

I thought I ought to atleast check that the formula (5) is correct. Verified it manually for small n , then wrote some hasty code in **MATLAB** to verify for larger n .

Feeds in $n = 10$ Correct!

Feeds in $n = 25$ Correct!

<heart-racing>

Feeds in $n = 100$ <churning churning> Correct!

Nothing more to say! It is indeed, an identity.

All That Glitters Is Not Gold

This is the combinatorial formula we have been horsing around with:

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

Let's observe it for a bit. Each term in the summation is actually a non-negative integer even though one exponent and one base take negative values for some k . Further, all terms are positive except the term for $k = 1$. Strangely reminiscent of the binomial formula $\sum_{k=0}^n \binom{n}{k} y^k (n-y)^{n-k} = n^n$, but not quite - the y depends on the k in some funny way. I mused over this formula for quite a while, but a proof eluded me - combinatorial, induction, direct fiddling with the terms involved. One crazy idea I had was to come up with a distribution for a random variable Y which somehow exhibited

$$\mathbb{E}[Y^k (n-Y)^{n-k}] = (k-1)^k (n-k+1)^{n-k-1}, \quad \text{for } 0 \leq k \leq n ,$$

but that was about all of it. Finally, after a full twenty days, I had occasion to go through my calculations again and discovered the flaw, the one from (1) to (2).

Atonement

I remember the moment of shock. I immediately corrected the mistaken step to get

$$\begin{aligned}
 V_{n,r}(t) &= \int_0^t \int_{y_1}^{2t} \int_{y_2}^{3t} \dots \int_{y_{n-1}}^{nt} (y_n)^r dy_n dy_{n-1} \dots dy_2 dy_1 \\
 &= \int_0^t \int_{y_1}^{2t} \int_{y_2}^{3t} \dots \int_{y_{n-2}}^{(n-1)t} \frac{(nt)^{r+1} - (y_{n-1})^{r+1}}{r+1} dy_{n-1} \dots dy_2 dy_1 \\
 c_{n,r} t^{n+r} &= \frac{(nt)^{r+1}}{r+1} V_{n-1,0}(t) - \frac{1}{r+1} V_{n-1,r+1}(t) \\
 c_{n,r} &= \frac{n^{r+1}}{r+1} c_{n-1,0} - \frac{1}{r+1} c_{n-1,r+1} \\
 &= \frac{n^{r+1}}{r+1} c_{n-1,0} - \frac{(n-1)^{r+2}}{(r+1)(r+2)} c_{n-2,0} + \frac{1}{(r+1)(r+2)} c_{n-2,r+2} \quad (\text{using (6) here}) \\
 &= \sum_{k=1}^{n-1} (-1)^{k-1} \frac{r!(n-k+1)^{k+r}}{(k+r)!} \cdot c_{n-k,0} + (-1)^{n+1} \frac{1}{(r+1)(r+2) \dots (r+n-1)} c_{1,n+r-1} \\
 c_{n,r} &= \sum_{k=1}^n (-1)^{k-1} \frac{r!(n-k+1)^{k+r}}{(k+r)!} \cdot c_{n-k,0} \quad \left(\text{using } c_{1,n+r-1} = \frac{1}{n+r}, c_{0,0} = c_0 = 1 \right)
 \end{aligned} \tag{6}$$

which is in fact, the *correct* recurrence formula for $c_{n,r}$, (as opposed to (3)) and which for $r = 0$ gives

$$\sum_{k=0}^n (-1)^k \frac{(n-k+1)^k}{k!} \cdot c_{n-k} = 0 . \tag{7}$$

Completing the proof by induction requires us to verify this by plugging in $c_l = \frac{(l+1)^{l-1}}{l!}$ for each l , i.e. to verify

$$\sum_{k=0}^n (-1)^k \binom{n}{k} (n-k+1)^{n-1} = 0 ,$$

which is easily seen to be true by observing that $\sum_{k=0}^n (-1)^k \binom{n}{k} k^r = 0$ for $0 \leq r \leq n-1$. This completes the solution to the fun problem.

But What About That ‘Identity’?

I biked vehemently to the computer room next morning, wrote up some more **MATLAB** code to check if the formula (5) was correct for each n . The machine got back announcing that the formula was true for all n in 1 to 100 except for

$$\begin{aligned}
 n=19, 22, 31, 33, 34, 37, 40, 43, 44, 45, 46, 49, 50, 53, 54, 57, 58, 63, 65, 67, \\
 68, 73, 75, 77, 78, 79, 80, 81, 83, 85, 87, 88, 89, 90, 91, 92, 93, 94, 95, 99
 \end{aligned}$$

i.e. false for some 40 integers between 1 to 100 with 19 being the first counterexample. True for some, false for some - certainly, a very odd way to behave, for a combinatorial formula that is wrong!

Redemption

Summer soon was over and I got back to dear old IIT Bombay, my alma mater. Early November '07, I received an unexpected e-mail. Written by Prof. Ajit Diwan, IIT Bombay, it began: “*By accident, I found your blog and the identity you mentioned that you found.*” Wait! I hadn’t told you that during those intern days, I had written a corny blogpost detailing this experience of almost-hitting-something-very-nice. The e-mail continued: “*It is actually true and there is a simple combinatorial proof for it. (I did not know it earlier).*”

WOW! Apparently, MATLAB had incorrectly reported that the formula was false for some n due to problems with precision. To be fair, the numbers involved, like 19^{19} or atleast 99^{99} , are huge!

.... But Gold Always Glitters

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

Quoting Prof. Diwan’s e-mail to me:

The R.H.S. is just the number of functions from $\{1, 2, ..n\}$ to itself. The L.H.S counts these in a different way. For any such function f , let $k(f)$ denote the largest i , such that $|f^{-1}\{1, 2, \dots, i-1\}| \geq i$, and $k(f) = 0$ if there is no such i . The L.H.S. counts the number of functions f with $k(f) = k$, for $k = 0$ to n . The tricky part is the $k = 0$ case. In this case, the number of functions is $(n+1)^{n-1}$. Note that this is just the number of spanning trees in the complete graph with $n+1$ vertices. It is possible to show this by an explicit bijection. The case for $k > 0$ follows easily from this.

This is just beautiful. The case $k = 0$ in Prof. Diwan’s proof above gives the set of *parking functions* but “in reverse”. A theorem due to Konheim and Weiss [1966] states that the number of parking functions of length n is $(n+1)^{n-1}$. You can find a nice proof of this theorem here: www-math.mit.edu/~rstan/transparencies/parking2.ps

But How?

After receiving this e-mail, I looked up my notes again. You see the silly error in going from (1) to (2)? Right! Now, note first that all subsequent steps are correct. And guess what? This mistaken step is actually correct only for the special case when $r = 0$ which is all that we end up using it for! So, while (3) is false, its special case (4) is actually true!

Two different *correct* recurrence relations for c_n can be obtained using very similar approaches, namely (7) and (4):

$$\sum_{k=0}^n (-1)^k \frac{(n-k+1)^k}{k!} \cdot c_{n-k} = 0 \qquad \sum_{k=0}^n \frac{(k-1)^k}{k!} \cdot c_{n-k} = \frac{n^n}{n!}$$

Yet somehow, on plugging $c_l = \frac{(l+1)^{l-1}}{l!}$, the left one leads to an easy combinatorial identity, while the right one leads to something much more interesting.