Contents

PROBABILITY LIFESAVER: Matching Coefficients and the Theory of Runs

Steven J. Miller

June 12, 2019

Chapter 1

Analyzing Runs

This is a supplemental chapter to *The Probability Lifesaver*.

1.1 Matching Coefficients

Sometimes we can derive identities of binomial coefficients without differentiating – one common technique is matching coefficients. For example, consider

$$\sum_{k=0}^{n} \binom{n}{k}^2 \ = \ \sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k},$$

because $\binom{n}{k} = \binom{n}{n-k}$. Consider now the following sum

$$\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \cdot \binom{n}{n-k} x^{n-k} y^k,$$

as well as

$$(x+y)^n (x+y)^n.$$

Expanding the product gives

$$(x+y)^n (x+y)^n \ = \ (x+y)^{2n} \ = \ \sum_{j=0}^{2n} \binom{2n}{j} x^j y^{2n-j};$$

note the coefficient of $x^n y^n$ in this product is $\binom{2n}{n}$. The key observation is that (1.1) is just the $x^n y^n$ term of $(x+y)^{2n}$. This is because it can be interpreted as taking the $x^n y^n$ term of $(x+y)^n(x+y)^n$. How do we get an $x^n y^n$ term from multiplying $(x+y)^n$ with

$(x+y)^n$? Well, the two factors $(x+y)^n$ give terms like $\binom{n}{i}x^iy^{n-i}$ and $\binom{n}{j}x^jy^{n-j}$, which are then multiplied together. The only way we get an x^ny^n is when $j = n-i$, and we can do this for any $j \in \{0, 1, \dots, n\}$. Thus the x^ny^n term in $(x+y)^{2n}$ is

$$\binom{2n}{n}x^ny^n \ = \ \sum_{k=0}^{n}\binom{n}{k}x^ky^{n-k}\cdot\binom{n}{n-k}x^{n-k}y^k.$$

The proof is completed by taking $x = y = 1$.

The reason arguments like this work is because if we have two polynomials of finite degree in finitely many variables, then if they take on identical values for all values of the parameters then all the coefficients of the two polynomials are equal. This allowed us to take two expressions and equate the coefficients of terms. Without this observation, the equality of two polynomials (at all values of the parameters) would not imply the equality of the coefficients. For example, assume $x^2+2xy-7y = x^2+3xy-5y^2+y$ for all $x, y \in \mathbb{C}$ (of course these two polynomials are not always equal); however, if this *were* to happen, we would be in trouble as in the first we have $2xy$ and the second we have $3xy$. Thus while some terms (such as x^2) have the same coefficient, others do not.

Specifically, say $F(x,y)$ and $G(x,y)$ are two polynomials of finite degree with complex coefficients. Then if they are equal for all choices of $x, y \in \mathbb{C}$ we have $F(x,y) - G(x,y)$ is a polynomial of finite degree and it is zero for all $x, y \in \mathbb{C}$. It is an easy exercise to show this implies all the coefficients of $F(x,y) - G(x,y)$ are zero (i.e., all the coefficients of $F(x,y)$ equal those of $G(x,y)$). One way to see this is to choose fixed values of x. Say $x = a$. Except for finitely many choices of a, we would get $F(a,y) - G(a,y)$ is a finite degree polynomial and it has some non-zero coefficient but it vanishes for all $y \in \mathbb{C}$. This is absurd as a polynomial of degree d has at most d complex roots. We do not need to have x and y range over all of $\mathbb{C}$; it suffices to have them range over a large enough set, for example $|x|, |y| \le R$ for some $R > 0$.

The biggest difficulty in successfully applying arguments of this nature is figuring out what to compare the observed sum to. Here we needed to see that we should compare $\sum_{k=0}^{n}\binom{n}{k}^2$ to the coefficient of x^ny^n in $(x+y)^{2n}$. Writing $\binom{n}{k}$ as $\binom{n}{k}\cdot\binom{n}{n-k}$ suggests that we should compare it to a coefficient of $(x+y)^n(x+y)^n$.

1.2 The Alternating Strings Problem

Consider a string of n_1+n_2 coin tosses with n_1 heads and n_2 tails. There are $\binom{n_1+n_2}{n_2}$ ways to order the n_1 heads and n_2 tails. Assume all orderings are equally likely. Our goal is to eventually study the number of alternating strings of heads and tails. We start with a simpler problem, namely trying to figure out how many ways there are to arrange n_1 heads and n_2 tails and observe u runs (again, $HHTTHTTTH$ would have 5 runs and 4 alterations).

For example, let us say $n_1 = n_2 = 3$ and we want to have 3 runs. If we assume we start with a head we could have $HTTTHH$ or $HHTTTH$, and by symmetry if we start with a tail we could have $THHHTT$ or $TTHHHT$.

In general, we have

Theorem 1.2.1. *Let there be n_1 heads and n_2 tails, and assume each of the $\binom{n_1+n_2}{n_1}$ arrangements are equally likely. Let there be u runs of heads and tails. Then*

$$u = \begin{cases} 2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1} & \text{if } u = 2k \text{ for a positive integer } k \\ \binom{n_1-1}{k}\binom{n_2-1}{k-1} + \binom{n_1-1}{k-1}\binom{n_2-1}{k} & \text{if } u = 2k+1 \text{ for a positive integer } k. \end{cases}$$

Proof. We consider $u = 2k$ and leave the other case as an exercise. As there are an even number of runs, we must either begin with a head and end with a tail, or we must begin with a tail and end with a head. By symmetry, it is enough to consider just the case when we start with a head and then multiply by 2. The reason is if we have a sequence like $HHHTTHTTTHTHT$ we can reverse it and obtain a sequence that starts with a tail and ends with a head.

Let us assume we will start with a head and end with a tail. Consider a string of n_1 heads. If we partition it into k strings of heads, we can then put tails in after the partitions, and we will have $2k$ runs; however, we *must* put a partition after the final head, as we must end with a tail. Further, we cannot put a partition before the first head as we *must* start with a head. For example, if we partition $HHHHH$ by adding partitions $|$ to get $H|HHH|H|$, then we can add strings of tails after the partitions to get $HT\cdots THHHT\cdots THT\cdots T$ for a total of 6 runs. How many ways are there to partition n_1 heads into k groups *with a partition occurring after the final head and no partition allowed before the first head*? Note there are $n_1 + 1$ positions where we can put a partition (before the first head, after the first head, after the second head, . . . , after the last head); however, we shall see that two of these positions have their values forced.

We must choose the last place for one partition, we cannot choose the place before the first head, and then we must choose $k - 1$ of the remaining $n_1 - 1$ positions for the other partitions. Thus the number of ways to add k partitions when we must add a partition after the final head and we cannot add one before the first head is just $\binom{1}{1}\binom{1}{0}\binom{n_1-1}{k-1} = \binom{n_1-1}{k-1}$. A similar argument shows there are $\binom{n_2-1}{k-1}$ ways to partition n_2 tails into k groups, assuming we must have a partition before the first tail and we are not allowed to have a partition after the final tail.

We now intersperse the partitioned heads and tails. Consider any of the $\binom{n_1-1}{k-1}$ partitions of the n_1 heads and any of the $\binom{n_2-1}{k-1}$ partitions of the n_2 tails. Each such pair gives rise to a sequence of n_1 heads and n_2 tails with exactly $2k$ runs, and any such sequence corresponds to a unique pair. For example, say we have $H|HH|HHH|$ and $|TTTT|T|TT$; these unite to become $HTTTTHHTHH$.

Thus the number of partitions leading to $2k$ runs where the first coin is a head and the last is a tail is just $\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}$. By symmetry this is the same as the number of partitions where the first coin is a tail and the last is a head, which completes the proof of the theorem in the case of an even number of runs. □

Of course, in the arguments above $1 \leq k \leq \min(n_1, n_2)$; for other k the number of strings with $2k$ runs is zero.

1.3 Determining How Often There are an Even Number of Runs

By differentiating identities we determine how often there are an *even* number of runs when there are n_1 heads and n_2 tails and each of the $\binom{n_1+n_2}{n_1}$ strings are equally likely. A similar argument is applicable for the case when there are an odd number of runs; we concentrate here on the case of an even number to highlight the methods.

If $u = 2k$ is the number of runs, then we know the number of ways to have $2k$ runs is just

$$2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}.$$

Without loss of generality, for notational convenience let us assume $n_1 \geq n_2$, so k runs from 1 to n_2. Thus the number of strings with an even number of runs is just

$$\sum_{k=1}^{n_2-1} 2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1},$$

as there must be at least two runs (there is no way to have zero runs unless $n_1 = n_2 = 0$, which we shall assume we do not have). We first need to determine what this sum is, and then to determine the expected number of u (when $u = 2k$ is even) we will need to sum

$$\sum_{k=1}^{n_2-1} (2k) \cdot 2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}.$$

1.3.1 Determining the number of strings with $u = 2k$ runs

Consider the polynomial

$$(x_1+y_1)^{n_1-1}(x_2+y_2)^{n_2-1};$$

we shall see very shortly why this is a "natural" polynomial to examine. Using the Binomial Theorem (Theorem **??**) we have

$$\begin{aligned}
(x_1+y_1)^{n_1-1} &= \sum_{k_1=0}^{n_1-1}\binom{n_1-1}{k_1}x_1^{n_1-1-k_1}y_1^{k_1} = \sum_{k_1=1}^{n_1}\binom{n_1-1}{k_1-1}x_1^{n_1-k_1}y_1^{k_1-1}\\
(x_2+y_2)^{n_2-1} &= \sum_{k_2=0}^{n_2-1}\binom{n_2-1}{k_2}x_2^{k_2}y_2^{n_2-1-k_2} = \sum_{k_2=1}^{n_2}\binom{n_2-1}{k_2-1}x_2^{k_2-1}y_2^{n_2-k_2};
\end{aligned}$$

we will see later why it is convenient to have $x_1^{n_1-k_1}$ but $x_2^{k_2-1}$; we can write the binomial theorem this way as $\binom{m}{r} = \binom{m}{m-r}$. Therefore

$$(x_1+y_1)^{n_1-1}(x_2+y_2)^{n_2-1} = \left[\sum_{k_1=1}^{n_1}\binom{n_1-1}{k_1-1}x_1^{n_1-k_1}y_1^{k_1-1}\right]\cdot\left[\sum_{k_2=1}^{n_2}\binom{n_2-1}{k_2-1}x_2^{k_2-1}y_2^{n_2-k_2}\right].$$

Consider what happens if we set $x_1 = x_2 = x$ and $y_1 = y_2 = y$. Then the above becomes

$$\begin{aligned}(x+y)^{n_1+n_2-2} &= \left[\sum_{k_1=1}^{n_1} \binom{n_1-1}{k_1-1} x^{n_1-k_1} y^{k_1-1}\right] \cdot \left[\sum_{k_2=1}^{n_2} \binom{n_2-1}{k_2-1} x^{k_2-1} y^{n_2-k_2}\right] \\ &= \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} \binom{n_1-1}{k_1-1}\binom{n_2-1}{k_2-1} x^{n_1-1-k_1+k_2} y^{n_2-1+k_1-k_2}.\end{aligned}$$

Now we use the uniqueness of polynomial expansions and equate coefficients. Consider the $x^{n_1-1}y^{n_2-1}$ term in (1.3.1). There are two ways we can calculate it. Looking at the left hand side, we have $(x+y)^{n_1+n_2-2}$, and thus the term is just $\binom{n_1+n_2-2}{n_1-1}x^{n_1-1}y^{n_2-1}$. Looking at the right hand side we see the term we desire occurs when $k_1 = k_2$. We see now why we wrote $x_1^{n_1-k_1}$ and $x_2^{k_2-1}$; this made it easy to combine the terms. Denoting the common value of k_1 and k_2 by k we obtain

$$\binom{n_1+n_2-2}{n_1-1} x^{n_1-1}y^{n_2-1} = \sum_{k=1}^{n_2} \binom{n_1-1}{k-1}\binom{n_2-1}{k-1} x^{n_1-1}y^{n_2-1},$$

or cancelling the x's and the y's

$$\binom{n_1+n_2-2}{n_1-1} = \sum_{k=1}^{n_2} \binom{n_1-1}{k-1}\binom{n_2-1}{k-1}.$$

We have determined the sum in (1.3), the sum we needed to figure out how many different strings there are with n_1 heads, n_2 tails and $u = 2k$ runs! Namely, we have shown

Lemma 1.3.1. *The number of strings with n_1 heads, n_2 tails and $u = 2k$ runs is*

$$\sum_{k=1}^{n_2-1} 2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1} = 2\binom{n_1+n_2-2}{n_1-1}.$$

Some discussion is clearly in order as to how we knew we should consider $(x_1 + y_1)^{n_1-1}(x_2+y_2)^{n_2-1}$. This is the hardest step in all such proofs by matching or proofs by differentiating identities, namely figuring out *where* to start. The answer is usually suggested by trying to analyze the quantity being studied, looking for clues as to what series or products we should consider.

In this case, we knew that we had to eventually have products like $\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}$. How can we get such terms? Well, the $\binom{n_1-1}{k-1}$ are the coefficients when we expand $(A+B)^{n_1-1}$; we chose $A = x_1$ and $B = y_1$ to have some flexibility, and to distinguish these terms from the other factors. For simply counting the number of strings with $u = 2k$ runs this extra degree of freedom or flexibility was not needed; however, it will be crucial in trying to find the mean of u when u is even. Similarly the $\binom{n_2-1}{k-1}$ are the coefficients from expanding $(A + B)^{n_2-1}$, and we choose $A = x_2$ and $B = y_2$ for the same reasons as before. By setting $x_1 = x_2 = x$ and $y_1 = y_2 = y$ in the end we are arguing in a similar manner as in §1.1. This is a common and powerful technique, namely writing $(x + y)^{n+m}$ and $(x + y)^n(x + y)^m$ and then deducing identities for sums involving terms like $\binom{n}{r}\binom{m}{a+r}$ for a fixed a.

1.3.2 Determining the expected value of u for strings with $u = 2k$ runs

We now turn to the sum in (1.3), which gives the expected value of $u = 2k$; again, remember that we are only considering strings with n_1 heads, n_2 tails and an even number $u = 2k$ of runs. As by Lemma 1.3.1 there are $2\binom{n_1+n_2-2}{n_1-1}$ such strings and the number of strings with $2k$ runs is $2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}$, we need to determine

$$\frac{\sum_{k=1}^{n_2-1}(2k)\cdot 2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{2\binom{n_1+n_2-2}{n_1-1}} = 2\frac{\sum_{k=1}^{n_2-1} k\cdot\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2-2}{n_1-1}}.$$

We shall ignore the factor of $2\binom{n_1+n_2-2}{n_1-1}^{-1}$ for now and concentrate on evaluating

$$\sum_{k=1}^{n_2-1} k\cdot\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}.$$

Actually, it will be significantly easier to find, not the sum with k but the sum with $k-1$:

$$\sum_{k=1}^{n_2-1}(k-1)\cdot\binom{n_1-1}{k-1}\binom{n_2-1}{k-1};$$

clearly if we can evaluate this sum for $k-1$ then by adding 1 we can find the sum with k.

We have seen in §1.3.1 that the sum over k of $\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}$ can be obtained by looking at the $x^{n_1-1}y^{n_2-1}$ coefficient of $(x_1+y_1)^{n_1-1}(x_2+y_2)^{n_2-1}$ under $x_1 = x_2 = x$ and $y_1 = y_2 = y$. So, let us study again (1.3.1):

$$(x_1+y_1)^{n_1-1}(x_2+y_2)^{n_2-1} = \left[\sum_{k_1=1}^{n_1}\binom{n_1-1}{k_1-1}x_1^{n_1-k_1}y_1^{k_1-1}\right]\cdot\left[\sum_{k_2=1}^{n_2}\binom{n_2-1}{k_2-1}x_2^{k_2-1}y_2^{n_2-k_2}\right].$$

We will now see the advantage of having two different x's and two different y's. Let us take the derivative with respect to y_1 and then multiply by y_1. Thus we are applying the operator $y_1\frac{\partial}{\partial y_1}$; the advantage of multiplying by y_1 after differentiating by y_1 is that we do not change the degree of any of the terms. Applying $y_1\frac{\partial}{\partial y_1}$ to the left hand side of (1.3.1) gives

$$(n_1-1)y_1(x_1+y_1)^{n_1-2}(x_2+y_2)^{n_2-1},$$

because x_1, y_1, x_2 and y_2 are independent variables. When we apply $y_1\frac{\partial}{\partial y_1}$ to the right hand side of (1.3.1) we get

$$\left[\sum_{k_1=1}^{n_1}(k_1-1)\binom{n_1-1}{k_1-1}x_1^{n_1-k_1}y_1^{k_1-1}\right]\cdot\left[\sum_{k_2=1}^{n_2}\binom{n_2-1}{k_2-1}x_2^{k_2-1}y_2^{n_2-k_2}\right].$$

The above shows why it is easier to study $k-1$ rather than k: when we differentiate a factor of $k-1$ comes down, not k. We have thus shown

$$\begin{aligned}&(n_1-1)y_1(x_1+y_1)^{n_1-2}(x_2+y_2)^{n_2-1}\\ &\quad=\left[\sum_{k_1=1}^{n_1}(k_1-1)\binom{n_1-1}{k_1-1}x_1^{n_1-k_1}y_1^{k_1-1}\right]\cdot\left[\sum_{k_2=1}^{n_2}\binom{n_2-1}{k_2-1}x_2^{k_2-1}y_2^{n_2-k_2}\right].\end{aligned}$$

NOW we take $x_1=x_2=x$ and $y_1=y_2=y$ and obtain

Lemma 1.3.2.

$$\begin{aligned}&(n_1-1)y(x+y)^{n_1+n_2-3}\\ &\quad=\left[\sum_{k_1=1}^{n_1}(k_1-1)\binom{n_1-1}{k_1-1}x^{n_1-k_1}y^{k_1-1}\right]\cdot\left[\sum_{k_2=1}^{n_2}\binom{n_2-1}{k_2-1}x^{k_2-1}y^{n_2-k_2}\right].\end{aligned}$$

It is extremely important that we waited to set x_1 equal to x_2 and y_1 equal to y_2; if we had set them equal first and then differentiated, we would have two pieces (from when the operator hit the first sum and when it hit the second). The difficulty would be the first sum would bring down a factor of k_1-1 and the second a factor of n_2-k_2. With some book-keeping this could probably be made to work, but this is easier.

We now look at the $x^{n_1-1}y^{n_2-1}$ term of both sides of Lemma 1.3.2. First consider the left hand side. We have one factor of y automatically because of the y outside. There are $\binom{n_1+n_2-3}{n_1-1}$ ways to choose n_1-1 factors of $(x+y)^{n_1+n_2-3}$ to give x and n_2-2 factors to give y. Thus the coefficient of $x^{n_1-1}y^{n_2-1}$ on the left hand side is

$$(n_1-1)\binom{n_1+n_2-3}{n_1-1}.$$

We now determine the $x^{n_1-1}y^{n_2-1}$ term from the right hand side of Lemma 1.3.2. As before, this term arises from $k_1=k_2$. Denoting this common value by k we find the coefficient of the $x^{n_1-1}y^{n_2-1}$ term from the right hand side is

$$\sum_{k=1}^{n_2}(k-1)\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}.$$

As always, the proof is concluded by the uniqueness of the coefficients. By matching we obtain

Lemma 1.3.3.

$$\sum_{k=1}^{n_2}(k-1)\binom{n_1-1}{k-1}\binom{n_2-1}{k-1} \;=\; (n_1-1)\binom{n_1+n_2-3}{n_1-1}.$$

We can now determine the mean of $k-1$, or better yet $2(k-1)$. From this it is trivial to determine the mean of $2k$. Specifically

Lemma 1.3.4.

$$\frac{\sum_{k=1}^{n_2-1} 2(k-1)\cdot 2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{2\binom{n_1+n_2-2}{n_1-1}} \;=\; 2\frac{n_1n_2-n_1-n_2+1}{n_1+n_2-3}.$$

Proof. The denominator comes from Lemma 1.3.1, where we showed this is the number of strings with n_1 heads, n_2 tails and an even number of runs. We cancel two of the factors of 2 and are left with one factor of 2 in the numerator, and then use Lemma 1.3.3 to evaluate the numerator. The proof is completed by expanding out the binomial coefficients. Let $\mu_{u-2,\text{even}}$ denote the mean of two less than even u (in other words, the expected value of $2(k-1)$ when $u = 2k$). Then

$$\begin{aligned}
\mu_{u-2,\text{even}} &= \frac{2(n_1-1)\binom{n_1+n_2-3}{n_1-1}}{\binom{n_1+n_2-2}{n_1-1}} \\
&= 2(n_1-1)\binom{n_1+n_2-3}{n_1-1}\cdot\binom{n_1+n_2-2}{n_1-1}^{-1} \\
&= \frac{2(n_1-1)(n_1+n_2-3)!}{(n_1-1)!(n_2-2)!}\cdot\frac{(n_1-1)!(n_2-1)!}{(n_1+n_2-2)!} \\
&= \frac{2(n_1+n_2-3)!}{(n_1-2)!(n_2-2)!}\cdot\frac{(n_1-1)!(n_2-1)!}{(n_1+n_2-2)(n_1+n_2-3)!} \\
&= \frac{2(n_1-1)(n_2-1)}{n_1+n_2-2} \\
&= 2\frac{n_1n_2-n_1-n_2+1}{n_1+n_2-2}.
\end{aligned}$$

Note that as we write $(n_1+n_2-2)!$ as $(n_1+n_2-2)\cdot(n_1+n_2-3)!$, we are implicitly assuming that $n_1+n_2-2 \geq 1$. If this fails, i.e. if $n_1+n_2 \leq 2$, then the above algebra could be wrong and those cases should be investigated separately (though if interpreted properly, our formulas will still be correct in these cases). $\square$

By adding 2 we get the mean of $u = 2k$ for even u.

Theorem 1.3.5. *Assume we have n_1 heads, n_2 tails, $u = 2k$ runs and all strings are equally likely. Then the expected number of runs is*

$$\mu_{u,\text{even}} \;=\; 2\left[\frac{n_1n_2-n_1-n_2+1}{n_1+n_2-2}+1\right] \;=\; 2\frac{n_1n_2-1}{n_1+n_2-2}.$$

Whenever one derives a complicated formula, it is a good idea to test it in extreme cases and see if it is reasonable. For example, the formula does not make sense if $n_1+n_2-2 = 0$. However, the only way that could happen, since n_1 and n_2 are non-negative integers, is if either both equal 1 or one is 0 and the other 2. If one is 0 and the other is 2 then we have an *odd* number of runs, and this formula is only for the case of an even number of runs. We are left with the case when $n_1 = n_2 = 1$. We have two runs, either HT or TH. In this case we have $2\frac{n_1n_2-1}{n_1+n_2-2} = 2\frac{0}{0}$; it is not unreasonable to think $\frac{0}{0}$ should be interpreted as 1 in this instance, and we would then get 2 (the correct answer). However, some care is needed in using this formula when $n_1 + n_2 = 2$, but this case can be handled directly.

Another good extreme to consider is when n_1 is much larger than n_2 (or vice-versa, but we have assumed without loss of generality earlier that $n_1 \geq n_2$). In this case, the mean for sequences with an even number of runs is approximately $2\frac{n_1 n_2}{n_1}$ or about $2n_2$. This is the correct behavior for such n_1 and n_2. Why? Imagine we have millions of time more heads (n_1) than tails (n_2). In that case it is extremely unlikely that any two tails will be adjacent. Thus there will be strings of varying lengths between the tails. As there are n_2 tails, this gives us $2n_2$ runs (the heads before a tail, a tail, another string of heads, a tail, another string of heads, a tail, and so on).

While such sanity checks are not proofs, they help us see if our formulas are reasonable, as well as possibly catching missing factors. For example, if we had dropped a factor of 2 earlier we would have found the mean was $\frac{n_1 n_2 - 1}{n_1 + n_2 - 2}$, and this would not have the right behavior for n_1 significantly larger than n_2. We also saw that the -2 in the denominator is reasonable.

We can also try a special case, for example $n_1 = 2$, $n_2 = 1$. In this case if we want an even number of runs we must have HHT or THH. Thus all strings with an even number of runs have 2 runs, and our formula does give 2 when $n_1 = 2$ and $n_2 = 1$. This helps check the -1 factor.

Thus, while it is still possible that we have made an algebra error somewhere, we should have a high degree of confidence in the result.

1.3.3 Determining the variance of u for strings with $u = 2k$ runs

Theorem 1.3.6. *Assume we have n_1 heads, n_2 tails, $u = 2k$ runs and all strings are equally likely. Then the variance in the number of runs is*

$$\sigma^2_{u,\text{even}} = 4\frac{(n_1-1)^2(n_2-1)^2}{(n_1+n_2-2)^2(n_1+n_2-3)}.$$

Proof. As $u = 2k$ is even, we need to find $\text{Var}(2k) = \mathbb{E}[(2k)^2] - \mathbb{E}[2k]^2$. We can simplify the calculations by noting that the variance of $u = 2k$ is the same as the variance of $u - 2 = 2(k-1)$. While we know the mean of both $u = 2k$ and $u - 2 = 2(k-1)$, it will turn out to be easier to calculate $\mathbb{E}[(2k-2)^2]$ than $\mathbb{E}[(2k)^2]$.

Thus we must evaluate

$$\frac{\sum_{k=1}^{n_2-1}\left[2(k-1)\right]^2 \cdot 2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{2\binom{n_1+n_2-2}{n_1-1}} = 4\frac{\sum_{k=1}^{n_2-1}(k-1)^2 \cdot \binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2-2}{n_1-1}}.$$

As before, the starting point is (1.3.1):

$$(x_1+y_1)^{n_1-1}(x_2+y_2)^{n_2-1} = \left[\sum_{k_1=1}^{n_1}\binom{n_1-1}{k_1-1}x_1^{n_1-k_1}y_1^{k_1-1}\right]\cdot\left[\sum_{k_2=1}^{n_2}\binom{n_2-1}{k_2-1}x_2^{k_2-1}y_2^{n_2-k_2}\right].$$

We apply the operator $x_2 y_1 \frac{\partial^2}{\partial x_2 \partial y_1}$. The reason for this choice is that the two derivatives bring down a factor of $(k_1-1)(k_2-1)$; the presence of $x_2 y_1$ means the degree of each term is unchanged (in all four variables x_1, x_2, y_1, y_2). Setting $x_1 = x_2 = x$ and $y_1 = y_2 = y$ and matching coefficients will complete the proof, as looking at the coefficient of $x^{n_1-1}y^{n_1-1}$ will cause $k_1 = k_2$, and this will give us the sum we desire.

Specifically, after applying $x_2y_1\frac{\partial^2}{\partial x_2 \partial y_1}$ the left hand side of (1.3.1) is

$$(n_1-1)(n_2-1)x_2y_1(x_1+y_1)^{n_1-2}(x_2+y_2)^{n_2-2},$$

while the right hand side of (1.3.1) is

$$\left[\sum_{k_1=1}^{n_1}(k_1-1)\binom{n_1-1}{k_1-1}x_1^{n_1-k_1}y_1^{k_1-1}\right]\cdot\left[\sum_{k_2=1}^{n_2}(k_2-1)\binom{n_2-1}{k_2-1}x_2^{k_2-1}y_2^{n_2-k_2}\right].$$

Setting $x_1 = x_2 = x$ and $y_1 = y_2 = y$, (1.3.3) and (1.3.3) give

$$\begin{aligned}(n_1-1)(n_2-1)xy(x+y)^{n_2+n_2-4} \quad &= \quad \left[\sum_{k_1=1}^{n_1}(k_1-1)\binom{n_1-1}{k_1-1}x^{n_1-k_1}y^{k_1-1}\right] \\ &\quad \cdot\left[\sum_{k_2=1}^{n_2}(k_2-1)\binom{n_2-1}{k_2-1}x^{k_2-1}y^{n_2-k_2}\right].\end{aligned}$$

We match the $x^{n_1-1}y^{n_1-1}$ term on both sides. The left hand side is easy. As we have an xy outside, we see we need to choose $n_1 - 2$ more x's and $n_2 - 2$ more y's. The right hand side is just the sum over $k_1 = k_2$. Denoting this common value by k we find

$$(n_1-1)(n_2-1)\binom{n_1+n_2-4}{n_1-2}x^{n_1-1}y^{n_2-1} \;=\; \sum_{k=1}^{n_2}(k-1)^2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}x^{n_1-1}y^{n_2-1},$$

or equivalently

$$\sum_{k=1}^{n_2}(k-1)^2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1} \;=\; (n_1-1)(n_2-1)\binom{n_1+n_2-4}{n_1-2}.$$

Therefore we have

$$\mathbb{E}[(2k-2)^2] \;=\; \frac{4(n_1-1)(n_2-1)\binom{n_1+n_2-4}{n_1-2}}{\binom{n_1+n_2-2}{n_1-1}}.$$

We can simplify the above expression to make it easier to subtract $\mathbb{E}[(2k-2)]^2$:

$$\begin{aligned}\mathbb{E}[(2k-2)^2] \quad &= \quad 4(n_1-1)(n_2-1)\frac{(n_1+n_2-4)!}{(n_1-2)!(n_2-2)!}\cdot\frac{(n_1-1)!(n_2-1)!}{(n_1+n_2-2)!} \\ &= \quad 4(n_1-1)(n_2-1)\frac{(n_1+n_2-4)!}{(n_1-2)!(n_2-2)!}\cdot\frac{(n_1-1)(n_1-2)!(n_2-1)(n_2-2)!}{(n_1+n_2-2)(n_1+n_2-3)(n_1+n_2-4)!} \\ &= \quad 4\frac{(n_1-1)^2(n_2-1)^2}{(n_1+n_2-2)(n_1+n_2-3)} \\ &= \quad 4\frac{(n_1-1)^2(n_2-1)^2}{(n_1+n_2-2)^2}\cdot\frac{n_1+n_2-2}{n_1+n_2-3}.\end{aligned}$$

We must now subtract $\mathbb{E}[(2k-2)]^2$. It is easiest algebraically to use the expression for $\mathbb{E}[(2k-2)]^2$ from the second to last line of (1.3.1). This yields

$$\begin{aligned}\text{Var}(2k-2) &= 4\frac{(n_1-1)^2(n_2-1)^2}{(n_1+n_2-2)^2}\cdot\frac{n_1+n_2-2}{n_1+n_2-3}-\left[\frac{2(n_1-1)(n_2-1)}{n_1+n_2-2}\right]^2 \\ &= 4\frac{(n_1-1)^2(n_2-1)^2}{(n_1+n_2-2)^2}\left[\frac{n_1+n_2-2}{n_1+n_2-3}-1\right] \\ &= 4\frac{(n_1-1)^2(n_2-1)^2}{(n_1+n_2-2)^2}\cdot\frac{1}{n_1+n_2-3} \\ &= 4\frac{(n_1-1)^2(n_2-1)^2}{(n_1+n_2-2)^2(n_1+n_2-3)},\end{aligned}$$

and $\text{Var}(2k-2)=\text{Var}(2k)$. □

For large n_1 and n_2,

$$\text{Var}(2k) \sim 4\frac{n_1^2n_2^2}{(n_1+n_2)^3}.$$

If n_1 is much larger than n_2, the mean is approximately $2n_2$ and the variance is approximately $4\frac{n_2^2}{n_1}$.

1.3.4 Behavior for all u

We briefly describe what happens if we don't restrict to the case when u, the number of runs, is even. The main result is that

Theorem 1.3.7. *Assume we have n_1 heads, n_2 tails, u runs and all strings are equally likely; u may be either even or odd and we assume $n_1, n_2 \geq 1$. Then the expected number of runs u is $\frac{2n_1n_2}{n_1+n_2}+1$ and the variance is $\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$. For n_1 and n_2 large, the expected number of runs is approximately $2\frac{n_1n_2}{n_1+n_2}$ and the variance is approximately $4\frac{n_1^2n_2^2}{(n_1+n_2)^3}$.*

Note our results on the expected number and variance of u (when u is forced to be even) are consistent with the above, at least when n_1 and n_2 are large. This isn't surprising, as when n_1 and n_2 are large it is reasonable to think that there are about as many strings with an odd number of runs as an even number of runs.

Sketch of the proof. To prove Theorem 1.3.7 we would need to investigate the case when $u=2k+1$. The starting point is the second part of (1.2.1), which tells us how many ways there are to have $u=2k+1$ runs. We need to know how many strings there are with n_1 heads and n_2 tails so that we can find the probabilities of having $u=2k$ or $u=2k+1$ runs. This is just $\binom{n_1+n_2}{n_1}$ as we choose n_1 of the n_1+n_2 positions to be heads.

In determining the mean and variance when $u=2k-2$ we divided the number of strings with $2k$ runs by $2\binom{n_1+n_2-2}{n_1-1}$, which is the number of strings with n_1 heads, n_2

tails and an even number of runs. What we can do is multiply our results on the mean and variance in this case by

$$\frac{2\binom{n_1+n_2-2}{n_1-1}}{\binom{n_1+n_2}{n_1}},$$

which now divides the contribution by the total number of strings and not just the total number of strings with an even number of runs.

The proof is completed by determining the contributions to the mean and the variance from the $u = 2k + 1$ terms. These contributions are found in a similar manner (i.e. by differentiating identities) as the $u = 2k$ terms. We leave the details to the reader. □

For completeness, we sketch the key steps in the algebra to finish the proof. We need to find the mean. For the terms with an even number of runs we need to average $2k$ and for the terms with an odd number of runs we average $2k + 1$.

For the even terms, we showed that there are $2\binom{n_1+n_2-2}{n-1}$ strings, and there are $\binom{n_1+n_2}{n_1}$ total strings. We multiply the mean in Theorem 1.3.5 by $\frac{2\binom{n_1+n_2-2}{n-1}}{\binom{n_1+n_2}{n_2}}$.

For the odd terms, from (1.2.1) we have two sums to study. To analyze the contribution from

$$\sum_k \binom{n_1-1}{k}\binom{n_2-1}{k-1}$$

we see this can be interpreted by looking at the $x^{n_2-2}y^{n_2}$ term of

$$\sum_{k_1} \binom{n_1-1}{k_1} x_1^{n_1-1-k_1} y_1^{k_1} \sum_{k_2} \binom{n_2-1}{k-1} x_2^{k_2-1} y_2^{n_2-k_2}$$

when we set $x_1 = x_2 = x$ and $y_1 = y_2 = y$. We see this term is the $x^{n_2-2}y^{n_2}$ term of

$$(x_1 + y_1)^{n_1-1}(x_2 + y_2)^{n_2-1}$$

when we set $x_1 = x_2 = x$ and $y_1 = y_2 = y$, and that term is just $\binom{n_1+n_2-2}{n_1-2}x^{n_1-2}y^{n_2}$. Note this allows us to determine the sum of these binomial coefficients. We need to evaluate the sum with a factor of $2k + 1$. To evaluate the sum with a factor of k we apply the operator $y_1 \frac{\partial}{\partial y_1}$; to handle the $+1$ in $2k+1$ we just need to count the number of terms, which from above is $\binom{n_1+n_2-2}{n_1-2}$. Therefore, the contribution from these terms with odd u from (1.2.1) to the mean is just

$$2(n_1-1)\binom{n_1+n_2-3}{n_1-2} + \binom{n_1+n_2-2}{n_1-2}$$

while the other terms with odd u in (1.2.1) give (by a similar argument or by symmetry) a contribution of

$$\binom{n_1+n_2-3}{n_2-2} + \binom{n_1+n_2-2}{n_2-2}.$$

We then must go through a lot of algebra - after adding all of these contributions we divide by the number of strings, $\binom{n_1+n_2}{n_1}$. In adding the various terms it is often convenient to pull out factors of $\frac{(n_1+n_2-3)!}{(n_1-2)!(n_2-2)!}$. In the end we show the mean is $\frac{2n_1n_2}{n_1+n_2}+1$. It is convenient to notice that

$$(n_1+n_2)(n_1+n_2-1)(n_1+n_2-2) \ = \ n_1^3+n_2^3+3n_1^2n_2+3n_1n_2^2-3n_1^2-3n_2^2-6n_1n_2+2n_1+2n_2.$$

Exercise 1.3.8. *Calculate the contributions from the $u = 2k+1$ terms and rescale the contributions from the $u = 2k$ terms to complete the proof of Theorem 1.3.7.*

1.3.5 Arbitrary Numbers of Heads and Tails

So far we assumed that there were n_1 heads, n_2 tails and all strings were equally likely. Let us assume now that we have N coin tosses where each toss has probability p of being a head and $q = 1-p$ of being a tail. Thus, $n_2 = N - n_1$. For each n_1 there are $\binom{N}{n_1}$ strings; all of these strings are equally likely, each occurring with probability $p^{n_1}q^{N-n_1}$. Our main result is

Theorem 1.3.9. *Assume we toss a coin with probability p of heads a total of N times. The expected number of runs $\mu_u(p)$ is $2p(1-p)(N-1)+1$. In particular, if the coin is fair (so $p = q = \frac{1}{2}$) then the expected number of runs is $\frac{N+1}{2}$.*

Proof. If there are n_1 heads then the expected number of runs is $\frac{2n_1(N-n_1)}{N}+1$, and there are $\binom{N}{n_1}$ such strings, each occurring with probability $p^{n_1}q^{N-n_1}$. Thus the expected number of runs $\mu_u(p)$ is

$$\begin{aligned}
\mu_u \quad &= \quad \sum_{n_1=0}^{N}\left[\frac{2n_1(N-n_1)}{N}+1\right]\cdot\binom{N}{n_1}p^{n_1}q^{N-n_1} \\
&= \quad 2\sum_{n_1=0}^{N}\frac{n_1(N-n_1)}{N}\frac{N!}{n_1!(N-n_1)!}p^{n_1}q^{N-n_1}+\sum_{n_1=0}^{N}\binom{N}{n_1}p^{n_1}q^{N-n_1} \\
&= \quad 2pq\sum_{n_1=1}^{N-1}\frac{(N-1)!}{(n_!-1)!(N-n_1-1)!}p^{n_1-1}q^{N-n_1-1}+(p+q)^N \\
&= \quad 2pq(N-1)\sum_{n_1=1}^{N-1}\frac{(N-2)!}{(n_1-1)!(N-n_1-1)!}p^{n_1-1}q^{N-n_1-1}+(p+q)^N \\
&= \quad 2p(1-p)(N-1)(p+q)^{N-2}+(p+q)^N.
\end{aligned}$$

As $q = 1-p$ the above becomes

$$\mu_u(p) = 2pq(N-1)+1.$$

In the special case that $p = q = \frac{1}{2}$ we have

$$\mu_u\left(\frac{1}{2}\right) \ = \ \frac{N+1}{2}.$$

□

Exercise 1.3.10. *Calculate the variance of* $\mu_u(p)$.