A Fortune Cookie Problem: A test for nominal data whether two samples are from the same population of equally likely elements (DRAFT)

Jiangtao Gou, Karen Ruth, Stanley Basickes and Samuel Litwin

April 4, 2020

Abstract

This article considers a way to test the hypothesis that two collections of objects are from the same uniform distribution of such objects. The exact p-value is calculated based on the distribution for the observed overlaps. In addition, an interval estimate of the number of distinct objects, when all objects are equally likely, is indicated.

KEYWORDS: recursion, hypergeometric, sampling with replacement, coupon collector problem.

1 Fortune cookie problem

The statisticians were sharing a batch of Dove$^{®}$ Promises$^{®}$ chocolate drops at a weekly meeting. Each candy wrapper contained a "fortune". For example, "Love is near", "Because you can" or the like. After reading many Dove$^{®}$ chocolate wrapper quotes, a question arose in this group of statisticians: Are all quotes from the same reservoir? Alternatively, does each kind of chocolate has some specific quotes which are not shared with other kinds of chocolates? For example, we would like to know whether the red-wrapped dark chocolate and blue-wrapped milk chocolate pick quotes from one collection, or from two collections which are not completely the same.

We can rephrase this problem with a setting of the coupon collector's problem (Dawkins 1991). Suppose each package of bubble gum contains a hockey card, and there are n different types of cards. We may ask whether the distribution of cards with bubble gum purchased in Pennsylvania and that of cards with bubble gum purchased in Illinois are the same. Equivalently, we would like to know if there is a type of card that is easier to be collected in

Illinois than in Pennsylvania, or vice versa. In general, we want to know if two nominal data sets were sampled from the same or different populations each of equally likely objects. We note that Wald and Wolfowitz (1940) proposed a test for continuous variables for whether two samples are from the same population using U-statistics. However, in order to apply U-statistics on nominal data sets, it is necessary to build an ordering of the factor levels of the variable. This article considers specifically the nominal data, so building an artificial ordering becomes unnecessary. One advantage of the proposed method is that only minimal information is required: the number of sampled elements and the number of the unique ones. If additional information is available, for example, a frequency table, we can test the assumption of discrete uniformity. For instance, a chi-squared goodness-of-fit test for discrete uniform distribution can be applied on Table 3, resulting p-values 0.983 for Bag One and 0.919 for Bag Two, which support our assumption of uniformity to a certain extent.

We first considered an estimation problem: How many distinct quotes does a collection have? For example, during one meeting, fortunes were recorded from 27 chocolates but we found only 17 different ones in a bag of dark chocolates. We wondered how many distinct fortunes we could find if we had an inexhaustable supply. Supposing all fortunes are equally likely and these $r = 27$ fortunes were selected at random, with replacement (Alf and Lohr 2007), from a total of n fortunes we wrote a recursion for the chance of finding k distinct fortunes after r chocolates have been examined:

$$p(k|r,n) = p(k|r-1,n) \cdot k/n + p(k-1|r-1,n) \cdot (n-k+1)/n,$$

where k is the number of distinct fortunes. Here $p(k|r-1,n)$ is the chance that k distinct fortunes have been seen up to the previous chocolate, and, assuming uniformity, k/n is the chance that the next selection will be the same as one already seen. Similarly, the term $p(k-1|r-1,n)$ is the chance that $k-1$ distinct fortunes have been seen and $(n-k+1)/n$ is the chance that the next selection will be a new one. The recursion is solved by:

$$p(k|r,n) = \left\{ {r \atop k} \right\} \cdot \frac{(n)_k}{n^r}, \tag{1}$$

where $(n)_k$ are the falling factorials as k-permutations of n, and $\left\{ {r \atop k} \right\}$ are Stirlings numbers of the second kind, which satisfy $\left\{ {r \atop k} \right\} = k \cdot \left\{ {r-1 \atop k} \right\} + \left\{ {r-1 \atop k-1} \right\}$ (Riordan 2002). We did discover this recursion and its solution in equation (1). To the best of our knowledge, they have not yet been described in the literature (Grimmett and Stirzaker 2001; Quaintance and Gould 2015). The interested reader need only express the recursion for a few increasing values of n and Stirlings numbers show up. The result then follows by induction. We illustrate this process in appendix 1, along with an explicit formula for Stirling number of the second kind.

We used this solution to find the maximum likelihood estimate, 26, and 95% confidence bounds, $[19, 41]$, of n. Our estimation method is related to the exact method of Clopper and Pearson (Clopper and Pearson 1934). They find bounds on the binomial parameter, p, from the number of successes, k in n Bernoulli trials. In our estimate of the number of distinct objects, n replaces p, the number k of distinct objects observed replaces their number, k of successes and the number of objects selected, r replaces their n. The two processes are parallel except that our n is discrete where their p is continuous, and our distribution is presented in equation (1) where Clopper and Pearson's distribution is binomial.

Our $(1-\alpha)\%$ confidence interval can be written as $[n_L, n_U]$, where

$$\Pr(k > k_0 \mid r = r_0, n = n_L) \le \alpha/2, \quad \Pr(k < k_0 \mid r = r_0, n = n_U) \le \alpha/2, \tag{2}$$

by equally allocating the significance level α to two sides. The upper value of the 95% confidence bound for n is the smallest n for which $\sum_{j=1}^{k-1} p(j|r,n) \le 0.025$, and the lower value is the largest n for which $\sum_{j=k+1}^{r} p(j|r,n) \le 0.025$. Plugging $k_0 = 17$ and $r_0 = 27$ into equation (2), we achieve $n_L = 19$ and $n_U = 41$. Moreover, without the assumption of equal allocation of the level of significance, we can further narrow the confidence interval in equation (2) to that in equation (3), where

$$\Pr(k > k_0 \mid r = r_0, n = n_L) = \alpha_L, \quad \Pr(k < k_0 \mid r = r_0, n = n_U) = \alpha_U, \tag{3}$$

and $\alpha_L + \alpha_U \le \alpha$. Using the observed $k_0 = 17$ and $r_0 = 27$, a list of 95% confidence intervals are presented in Table 1, where the least interval length is 20.

For the estimation problem, two equivalent descriptions of this problem are:

Table 1: Nominal 95% confidence intervals for n with $k = 17$ and $r = 27$

Method	C.I.	MLE	α_L	α_U	$\alpha_L + \alpha_U$	Interval length
Eq (2)	$[19, 41]$	26	0.0127	0.0233	0.0359	22
Eq (3)	$[17, 37]$	26	0.0000	0.0487	0.0487	20
	$[18, 38]$	26	0.0027	0.0404	0.0431	20
	$[19, 39]$	26	0.0127	0.0336	0.0462	20
	$[20, 44]$	26	0.0340	0.0136	0.0476	24

a. Selecting r balls at random from an urn containing n equally likely and uniquely identified balls, noting each selected balls' identification, then replacing it before the next selection is made. k different balls are identified. There is no concern over which balls are identified. And

b. Throwing r balls at random into n equally likely distinct cells, k cells become occupied. There is no concern over which cells become occupied.

Before getting to the two-sample comparison problem, we conclude the estimation problem with a simulation study. The likelihood function is provided in equation (1) and the confidence intervals are provided in equation (2) and (3). A simulation study is performed as follows. We randomly sample r observations with replacement from a factor with n levels, where each level has the same chance to be chosen. Next we find the number of the unique elements k and calculate the maximum likelihood estimate $\widehat{n}_{\text{MLE}}$. We then compute the 95% confidence interval using equation (3) with minimum interval length, which often results $\alpha_L < \alpha_U$ as defined in equation (3). We check if n falls in the interval $[\widehat{n}_L, \widehat{n}_U]$ which is estimated based on the prespecified r and the observed k. Here we report the results for $n = 20$ and 40. For $n = 20$, we consider $r = 10$, 20 and 30; and for $n = 40$, we consider $r = 20$, 40 and 60. The number of replica for each simulation is 10^4. We also report the average and standard deviation of $\widehat{n}_{\text{MLE}}$, as shown in Table 2. We observe that the simulated confidence intervals can be anti-conservative, where the nominal confidence level can be fell short by less than 2%.

Table 2: Simulated confidence levels for the confidence intervals proposed in equation (3)

n	r	mean($\widehat{n}_{\text{MLE}}$)	SD($\widehat{n}_{\text{MLE}}$)	$\widehat{\Pr}(n \in [\widehat{n}_L, \widehat{n}_U])$
$n = 20$	$r = 10$	22.2	12.5	0.934
	$r = 20$	20.8	6.2	0.957
	$r = 30$	19.8	3.2	0.943
$n = 40$	$r = 20$	47.5	28.8	0.942
	$r = 40$	40.4	8.1	0.938
	$r = 60$	39.8	4.5	0.954

At a subsequent meeting (in addition to business) we collected similar data on another batch of $r_2 = 30$ chocolates in a bag of milk chocolates. We found 18 distinct fortunes, 12 of which were also present in the first batch. Our problem then, was to decide if the two data sets were sampled from the same or different fortune collections.

A discussion ensued about the distribution of the overlap. We came to the idea that if the two samples came from the same batch then the overlap was governed by the hypergeometric distribution (Dwass 1979). Namely,

$$p(j|k_1, k_2, r_2, n) = \binom{k_1}{j}\binom{n-k_1}{k_2-j}\Big/\binom{n}{k_2} \tag{4}$$

where j fortunes are common to both selections, k_1 distinct fortunes were present in the first sample, k_2 distinct in the second and r_2 fortunes were collected in the second data set. We considered k_1 and k_2 fixed. Our intuition told us that the conditional probability $p(j|k_1, k_2, r_2, n)$ does not depend on r_2 directly. Following this intuition, we arrived at the hypergeometric distribution. A complete proof can be found in appendix 2 and we verified our intuition.

We struggled trying to prove that this was the case and that, as long as r_2 was at least k_2 the result should not involve r_2. One argument was: Since all r_2 balls are selected at random, each ball is equally likely to be seen at each draw. It seems easy to conclude that all subsets of k_2 balls are equally likely to have been drawn. This is the same as choosing all k_2 at once, without replacement. This implies that the distribution of overlap is hypergeometric. Simulation supported this conclusion.

We were pretty sure this was correct, but not convinced by the above argument. We sought a more direct path, and found:

$$p(j, k_2|k_1, r_2, n) = p(j, k_2|k_1, r_2 - 1, n) \cdot \frac{k_2}{n} + p(j-1, k_2-1|k_1, r_2-1, n) \cdot \frac{k_1 - j + 1}{n} + p(j, k_2 - 1|k_1, r_2 - 1, n) \cdot \frac{n - k_1 - k_2 + j + 1}{n}, \tag{5}$$

where each term relates the situation after $r_2 - 1$ selections followed by the chance that the next selection will lead to exactly k_2 distinct selections with j of them common to k_1 and k_2. Furthermore, the three terms comprise all possible paths to j overlaps with k_2 distinct selections when exactly one new selection is to be made.

Having $p(0, 1|k_1, 1, n) = (n-k_1)/n$ and $p(1, 1|k_1, 1, n) = k_1/n$ as initial conditions allowed us to compute the joint distribution $p(j, k_2|k_1, r_2, n)$. We summed this distribution over terms with a particular fixed k_2 and normalized these terms to one to get the conditional distribution $p(j|k_1, k_2, r_2, n)$. We found it to be equal to the hypergeometric with the same parameters, except for r_2.

We postulated that the joint distribution:

$$p(j, k_2|k_1, n, r_2) = \text{hypergeometric} \cdot p(k_2|r_2, n) = \binom{k_1}{j}\binom{n-k_1}{k_2-j}\Big/\binom{n}{k_2} \cdot p(k_2|r_2, n),$$

where the probability $p(k_2|r_2, n)$ can be found in (1). We found (see appendix 2) that this expression satisfies the recursion. Accepting this as the solution to the joint distribution problem, the hypergeometric is then seen to be:

$$\text{hypergeometric} = p(j, k_2|k_1, n, r_2)/p(k_2|r_2, n) = p(j|k_1, k_2, n, r_2)$$

establishing the result we believed to be true.

The null hypothesis that both samples are from the same batch can be tested by asking how likely is the overlap to be as small as or smaller than that observed. We were now confident that we could use the hypergeometric for this purpose. 95% confidence bounds on the total number of fortunes, n, as noted above are $[19, 41]$ from the first batch. Similarly they are $[20, 39]$ in the second batch which had 30 fortunes, 18 of which are distinct, with the assumption of equal allocation of α in equation (2). We considered all n in the range

$[19, 41]$. The hypergeometric distribution for the overlap, j, is

$$p(j|k_1 = 17, k_2 = 18, n) = \binom{17}{j}\binom{n-17}{18-j}\Big/\binom{n}{18}.$$

For an overlap, j, to be possible we must have $18 - j \leq n - 17$ or $35 - j \leq n$. We observed $j = 12$ from these two data sets, so $35 - 12 = 23 \leq n$. Hence we tested all n in $[23, 41]$. For each n in this range we computed the sum of $p(j)$ over all j in $[\max\{0, n - 35\}, 12]$. The smallest of these was 0.1839 obtained for $n = 23$. This probability is 0.4138 for $n = 24$ and increases rapidly to over 0.9 for $n > 28$. Hence, acceptance of the null hypothesis is justified.

In a general way, we calculate the probability $\Pr(j \leq j_0|k_1, k_2, r_1, r_2)$ as a measure of significance, where j_0 is the observed overlap. This tail-area probability can be treated as the p-value (Rubin 1984; Berger and Boos 1994; Meng 1994; Schervish 1996). We have proved that $p(j|k_1, k_2, r_1, r_2) = p(j|k_1, k_2)$ and showed that $p(j|k_1, k_2, n)$ follows a hypergeometric distribution. Therefore, $\Pr(j \leq j_0|k_1, k_2, r_1, r_2)$ can be calculated as $\sum_{\text{all possible } n} \Pr(j \leq j_0|k_1, k_2, n) \cdot p(n|k_1, k_2, r_1, r_2)$. Let $k = k_1 + k_2 - j$ and $r = r_1 + r_2$. We find $p(n|k_1, k_2, r_1, r_2)$ using $p(n|k, r) \propto p(k|r, n) \cdot p(n)$, where $p(n)$ is a prior distribution of n. In this article, we assume $p(n)$ follows a discrete uniform distribution on the integers in the 95% confidence interval calculated using equation (2) or (3). In our fortune cookie problem, a total of $r_1 + r_2 = 57$ chocolates were recorded and $k_1 + k_2 - j = 23$ difference ones were found. Using equation (3), a 95% confidence interval of n is computed as $[23, 29]$. Using the discrete uniform distribution on $\{23, \ldots, 29\}$ as a prior, we can find the values of $p(n|k_1, k_2, r_1, r_2)$. In addition, the probability $\Pr(j \leq j_0|k_1, k_2, n)$ can be evaluated using the hypergeometric distribution. For example, when $n = 23$, the probabilities $p(n|k_1, k_2, r_1, r_2) = 0.0821$ and $\Pr(j \leq j_0|k_1, k_2, n) = 0.1839$; when $n = 29$, the probabilities $p(n|k_1, k_2, r_1, r_2) = 0.0712$ and $\Pr(j \leq j_0|k_1, k_2, n) = 0.9349$. Combing these probabilities, we achieve $\Pr(j \leq j_0|k_1, k_2, r_1, r_2) = 0.6565$, and we tend to accept the null hypothesis.

The question we are interested here in this paper is whether two collections of items are from the same uniform distribution of such items. If we want to test the homogeneity without the assumption of the uniform distribution, we can apply other methods with additional information: the count data provided in appendix 3. For the first two bags of chocolates, we can view it as a 23 by 2 contingency table. Given the large number of cells with small expected counts, a permutation test is preferred (Ludbrook and Dudley 1998; David 2008; Zhang and Chen 2018). A permutation version of the chi-square test is performed with a number of samples 10^6, we get the simulated p-value 0.7066.

2 Discussion

We first estimated the total number of equally likely distinct fortunes, n, when r_1 samples yield k_1 distinct ones. Next, a second sample of size r_2 yielded k_2 distinct ones j of which had been seen in the first sample. We find a recursion for the joint distribution of j and k_2 given r_2, k_1 and n and show that this distribution is the product of a hypergeometric distribution and the distribution governing k_1 given r_1 and n. This result establishes the hypergeometric as the distribution of overlap j under the null hypothesis. The hypergeometric is then used to test the hypothesis that the two samples came from the same batch.

In the following meetings, two more batches of chocolates were consumed (bags 3 and 4 in the data set in appendix 3). With full information, this 40 by 4 contingency table, we performed a permutation chi-square test as a test of homogeneity, and got the simulated p-value 0.001. A question, then, arises: With only limited information available, how can we test the null hypothesis that four data sets are from the same collection, if we only have binary data? Several additional questions are suggested: Are bags 3 and 4 from the same batch as 1 and 2? Are bags 3 and 4 from the same batch as each other? Is the distribution of the multiplets what would be expected from random sampling? Is bag 3 with one fortune represented 5 times an outlier? How to apply the multiple testing procedures for correcting multiplicity (Hochberg and Tamhane 1987; Shaffer 1995; Westfall and Wolfinger 1997; Tamhane and Gou 2018)? We leave the interested reader with this data set to ponder.

Appendix

Appendix 1. First, we give a brief introduction of Stirling numbers of the second kind. A Stirling number of the second kind $\left\{ {r \atop k} \right\}$ describes how many ways to partition a set of r items into k subsets. An explicit formula for Stirling number of the second kind is

$$\left\{ {r \atop k} \right\} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^r ,$$

which is known as Euler's formula (Quaintance and Gould 2015). This formula perhaps is not as famous as other Euler's formulae. For fixed r, $\left\{ {r \atop k} \right\}$ first increases then decreases as a function of k, and the maximum is achieved near $k \sim r/\log r$. In addition, an asymptotic form of the Stirling numbers of the second kind when r goes to infinity is $\left\{ {r \atop k} \right\} \sim k^r/k!$. An example of comparisons among some combinatorial numbers is: $\binom{100}{30} \approx 10^{25}$, $(100)_{30} \approx 10^{58}$, $\left\{ {100 \atop 30} \right\} \approx 10^{115}$ and $100! \approx 10^{158}$.

Next, we prove equation (1). Starting with initial conditions, $p(k=1|r=2,n) = 1/n$, $p(k=2|r=2,n) = 1-(1/n)$ and $p(k=0|r=2,n) = 0$, if both r and n are positive, we find the probabilities by using the recursion successively. Proof then follows by induction: Suppose equation (1) holds for all $r \le r_0$ and all $k \le r_0$. We first consider $r = r_0+1$ and $k \le r_0$ and have

$$\begin{aligned} p(k|r_0+1,n) &= p(k|r_0,n) \cdot \frac{k}{n} + p(k-1|r_0,n) \cdot \frac{n-k+1}{n} \\ &= \left\{ {r_0 \atop k} \right\} \cdot \frac{(n)_k}{n^{r_0}} \cdot \frac{k}{n} + \left\{ {r_0 \atop k-1} \right\} \cdot \frac{(n)_{k-1}}{n^{r_0}} \cdot \frac{n-k+1}{n} \\ &= k \cdot \frac{(n)_k}{n^{r_0+1}} \cdot \left\{ {r_0 \atop k} \right\} + \frac{(n)_k}{n^{r_0+1}} \cdot \left\{ {r_0 \atop k-1} \right\} \\ &= \frac{(n)_k}{n^{r_0+1}} \cdot \left\{ {r_0+1 \atop k} \right\}. \end{aligned}$$

Next, for $k = r_0+1$, noting that $\left\{ {r \atop r} \right\} = 1$, we have

$$\begin{aligned} p(r_0+1|r_0+1,n) &= p(r_0|r_0,n) \cdot \frac{n-r_0}{n} \\ &= \left\{ {r_0 \atop r_0} \right\} \cdot \frac{(n)_{r_0}}{n^{r_0}} \cdot \frac{n-r_0}{n} = \left\{ {r_0+1 \atop r_0+1} \right\} \cdot \frac{(n)_{r_0+1}}{n^{r_0+1}}. \end{aligned}$$

Therefore, equation (1) holds for $r = r_0+1$ and all $k \le r_0+1$, completing the proof by induction.

Appendix 2. First, we will show that

$$p(j,k_2|k_1,n,r_2) = \left\{ {r_2 \atop k_2} \right\} \frac{(n)_{k_2}}{n^{r_2}} \cdot \frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2}} \tag{6}$$

is the solution of equation (5).

We note that $\binom{n}{k} = (n-k+1)/k \cdot \binom{n}{k-1}$. Hence

$$\begin{aligned}
&p(j-1,k_2-1|k_1,n,r_2-1)\cdot\frac{k_1-j+1}{n} + p(j,k_2-1|k_1,n,r_2-1)\cdot\frac{n-k_1-k_2+j+1}{n}\\
&= \frac{\binom{k_1}{j-1}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2-1}}\left\{ {r_2-1 \atop k_2-1} \right\}\frac{(n)_{k_2-1}}{n^{r_2-1}}\cdot\frac{k_1-j+1}{n}\\
&\quad + \frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j-1}}{\binom{n}{k_2-1}}\left\{ {r_2-1 \atop k_2-1} \right\}\frac{(n)_{k_2-1}}{n^{r_2-1}}\cdot\frac{n-k_1-k_2+j+1}{n}\\
&= \left\{ {r_2-1 \atop k_2-1} \right\}\frac{(n)_{k_2}}{n^{r_2}}\cdot\left[\frac{\binom{k_1}{j-1}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2-1}}\cdot\frac{k_1-j+1}{n-k_2+1} + \frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j-1}}{\binom{n}{k_2-1}}\cdot\frac{n-k_1-k_2+j+1}{n-k_2+1}\right]\\
&= \left\{ {r_2-1 \atop k_2-1} \right\}\frac{(n)_{k_2}}{n^{r_2}}\cdot\left[\frac{\binom{k_1}{j-1}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2}}\cdot\frac{k_1-j+1}{k_2} + \frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j-1}}{\binom{n}{k_2}}\cdot\frac{n-k_1-k_2+j+1}{k_2}\right]\\
&= \left\{ {r_2-1 \atop k_2-1} \right\}\frac{(n)_{k_2}}{n^{r_2}}\cdot\left[\frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2}}\cdot\frac{j}{k_2} + \frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2}}\cdot\frac{k_2-j}{k_2}\right]\\
&= \left\{ {r_2-1 \atop k_2-1} \right\}\frac{(n)_{k_2}}{n^{r_2}}\cdot\frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2}}.
\end{aligned}$$

Meanwhile, we have

$$p(j,k_2|k_1,n,r_2-1)\cdot\frac{k_2}{n} = k_2\cdot\left\{ {r_2-1 \atop k_2} \right\}\frac{(n)_{k_2}}{n^{r_2}}\cdot\frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2}}.$$

We use the fact that

$$k_2\left\{ {r_2-1 \atop k_2} \right\} + \left\{ {r_2-1 \atop k_2-1} \right\} = \left\{ {r_2 \atop k_2} \right\}.$$

As a result, we have

$$\begin{aligned}
&p(j,k_2|k_1,n,r_2-1)\cdot\frac{k_2}{n} + p(j-1,k_2-1|k_1,n,r_2-1)\cdot\frac{k_1-j+1}{n}\\
&\quad + p(j,k_2-1|k_1,n,r_2-1)\cdot\frac{n-k_1-k_2+j+1}{n}\\
&= k_2\cdot\left\{ {r_2-1 \atop k_2} \right\}\frac{(n)_{k_2}}{n^{r_2}}\cdot\frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2}} + \left\{ {r_2-1 \atop k_2-1} \right\}\frac{(n)_{k_2}}{n^{r_2}}\cdot\frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2}}\\
&= \left\{ {r_2 \atop k_2} \right\}\frac{(n)_{k_2}}{n^{r_2}}\cdot\frac{\binom{k_1}{j}\binom{n-k_1}{k_2-j}}{\binom{n}{k_2}}\\
&= p(j,k_2|k_1,n,r_2).
\end{aligned}$$

Second, note that the solution in (6) satisfies the initial conditions $p(j=0,k_2=1|k_1,n,r_2=$

$1) = (n - k_1)/n$ and $p(j = 1, k_2 = 1 | k_1, n, r_2 = 1) = k_1/n$. Therefore it is the only solution of equation (5).

Appendix 3. Our chocolate study data collection was informal. At our initial meeting, as we enjoyed our chocolates we shared our fortunes with each other and were surprised that there were duplicates in our relatively small sample. So, being statisticians, we saved the wrappers and started our own 'experiment'. On a simple tally sheet with each fortune on a separate line, we used hash marks to keep a running count. We left the remainder of the initial bag of chocolates on a counter with a request to add each fortune to the tally before throwing out the wrapper. We had started with an open bag, so we had results for only 27 of the 32 in the bag.

The person who had brought in the bag of chocolates had a second bag in her desk (purchased at the same time as the first), and so the experiment continued at our next meeting, this time using a red pen on the tally sheet. We left the remaining chocolates out on the honor system, asking for tallies. Two people must have been too rushed to make a mark — we ended up with an additional 30 chocolates.

After a while, we missed having chocolate at our meetings, and invested in another two bags of chocolates. We had bag ♯3 (28 chocolates tracked) and bag ♯4 (only 26 tracked – perhaps some survey fatigue was creeping in).

We copied the tally sheet into a spreadsheet, using the first letter of each word in the fortune as the identifier. The data are shown below.

References

Alf, C. and Lohr, S. (2007), "Sampling assumptions in introductory statistics classes," *The American Statistician*, 61, 71–77.

Berger, R. L. and Boos, D. D. (1994), "*P* Values Maximized Over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89, 1012–1016.

Table 3: Fortune data (40 types of fortunes)

	Bags of chocolate				
Fortune	Bag One	Bag Two	Bag Three	Bag Four	**Total**
identifier	Red	Blue	Green	Yellow	
ktg	3				**3**
suttsr	1	3	1		**5**
bb	1	1			**2**
tarotws	3		1	3	**7**
gm	2	1			**3**
stbff	2	3	2	3	**10**
ih	2	1			**3**
aac	3	4	2	2	**11**
gduwnptg	2	1			**3**
cdt	1	2		1	**4**
lsip	1	1			**2**
stotf	1				**1**
gtstta	1	1			**2**
dwwtnt	1				**1**
buc	1			1	**2**
lypb	1	2			**3**
tygttas	1	3			**4**
waywf		2	1	2	**5**
ttlf		1			**1**
mtfm		1	2	2	**5**
wi		1			**1**
bpoya		1			**1**
supyb		1			**1**
hub			1		**1**
cybfi			3		**3**
rtaoc			3		**3**
wn			5	1	**6**
gsac			1		**1**
rtlpf			1		**1**
lsnaalo			1		**1**
sawado			1		**1**
i			2		**2**
agtabs			1	1	**2**
itc				3	**3**
bml				2	**2**
hsx5				1	**1**
rwttd				1	**1**
hsa				1	**1**
sol				1	**1**
wttboyot				1	**1**
N recorded	27	30	28	26	**111**
N not recorded	5*	2	4	6	**17**

∗ started with open bag.

Clopper, C. J. and Pearson, E. S. (1934), "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, 26, 404–413.

David, H. A. (2008), "The Beginnings of Randomization Tests," *The American Statistician*, 62, 70–72.

Dawkins, B. (1991), "Siobhan's problem: The coupon collector revisited," *The American Statistician*, 45, 76–82.

Dwass, M. (1979), "A generalized Binomial distribution," *The American Statistician*, 33, 86–87.

Grimmett, G. R. and Stirzaker, D. R. (2001), *Probability and Random Processes*, Oxford, United Kingdom: Oxford University Press, 3rd ed.

Hochberg, Y. and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, New York, New York: John Wiley and Sons.

Ludbrook, J. and Dudley, H. (1998), "Why Permutation Tests are Superior to t and F Tests in Biomedical Research," *The American Statistician*, 52, 127–132.

Meng, X.-L. (1994), "Posterior Predictive p-Values," *The Annals of Statistics*, 22, 1142–1160.

Quaintance, J. and Gould, H. W. (2015), *Combinatorial Identities for Stirling Numbers*, Singapore: World Scientific Publishing.

Riordan, J. (2002), *Introduction to Combinatorial Analysis*, Mineola, New York: Dover Publications.

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.

Schervish, M. J. (1996), "P Values: What They Are and What They Are Not," *The American Statistician*, 50, 203–206.

Shaffer, J. P. (1995), "Multiple hypothesis testing," *Annual Review of Psychology*, 46, 561–584.

Tamhane, A. C. and Gou, J. (2018), "Advances in p-value based multiple test procedures," *Journal of Biopharmaceutical Statistics*, 28, 10–27.

Wald, A. and Wolfowitz, J. (1940), "On a test whether two samples are from the same population," *The Annals of Mathematical Statistics*, 11, 147–162.

Westfall, P. H. and Wolfinger, R. D. (1997), "Multiple tests with discrete distributions," *The American Statistician*, 51, 3–8.

Zhang, J. and Chen, C. (2018), "On "A mutual information estimator with exponentially decaying bias" by Zhang and Zheng," *Statistical Applications in Genetics and Molecular Biology*, 17.