Expected distinct values when selecting from a bag without replacement

Alberto Dell'Era, December 2005-August 2007, Version 1.1

Abstract

Derivation of the formula that gives the expected number of distinct values D_v contained in a sample of s balls, without replacement, from a bag (urn, box) containing N_r balls labeled with N_d distinct values (or distinct colors, etc).

1 Examples and definitions

1.1 Bag and Bucket

Consider a bag containing $N_r = 12$ balls and $N_d = 4$ distinct values $(0, 1, 2, 3)$:

0	0	0	0	0	$N_b(0) = 5$
1	1				$N_b(1) = 2$
2	2	2			$N_b(2) = 3$
3	3				$N_b(3) = 2$

We'll call the set of balls with the same label value a *bucket*, and define $N_b(k)$ as the "number of balls in the k-th bucket" (ie, the "number of balls with the same k-th value (color)").

The vector $N_b(k)$ defines the *distribution* of the values; when $N_b(k) = constant = N_b$ the distribution is *uniform*, otherwise it is *skewed*.

$N_b(k)$ can also be called a *frequency histogram*.

1.2 Samples and Distinct Values

Here's some examples of samples with $s = 5$ from our bag, and the resulting D_v :

{0 1 0 2 3}	$D_v = 4$
{2 1 2 1 3}	$D_v = 3$
{0 0 0 0 0}	$D_v = 1$

Considering all possible samples with $s = 5$ from our bag, we could see that the average (aka *expected value*) of D_v is $E[D_v] \approx 3.18$.

Actually:

s	$E[D_v]$
0	0
1	1
2	≈ 1.77
3	≈ 2.37
4	≈ 2.83
5	≈ 3.18
...	...
9	≈ 3.90
10	≈ 3.97
11	4.0
12	4.0

Note that for $s = 11$, no bucket can escape from the sampling hand.

2 Formulae

2.1 Skewed (general) distribution

$$\boxed{E[D_v] = \sum_{k=0}^{N_d-1} \left[1 - \prod_{i=0}^{N_b(k)-1} \left(1 - \frac{s}{N_r - i} \right) \right]} \quad (1)$$

Interesting bits:
$s = 0$ (no ball selected) $\Rightarrow E[D_v] = 0$
$s = 1$ (we select exactly one ball) $\Rightarrow E[D_v] = 1$
$s = N_r$ (we select all balls) $\Rightarrow E[D_v] = N_d$

2.2 Uniform distribution

For $N_b(i) = constant = N_b = N_r/N_d$, (1) becomes:

$$E[D_v] = N_d \left[1 - \prod_{i=0}^{N_r/N_d-1} \left(1 - \frac{s}{N_r - i} \right) \right] \quad (2)$$

Interesting bits:
$N_r = N_d$ (all values distinct) $\Rightarrow E[D_v] = s$
If we can ignore i for some reason (eg $i << N_r$), this formula can be approximated by:

$$E[D_v] = N_d \left(1 - \left(1 - \frac{s}{N_r} \right)^{N_r/N_d} \right)$$

2.3 Weakly uniform distribution

A "weakly uniform distribution" is the one obtained by adding zero or one ball to each bucket of a perfectly-uniform-distribution bag. This is obviously a special case of a "skewed" distribution, but it's important since it's the one normally assumed in practice when we don't know the actual distribution, and we know only N_r and N_d.

In this case, we have $(N_r \mod N_d) \equiv N_d^{bb}$ "big buckets" containing exactly one ball more than the remaining $(N_d - N_r \mod N_d) \equiv N_d^{sb}$ "small buckets". The number of rows contained in each bucket class is $ceil(N_r/N_d) \equiv N_b^{bb}$ and $floor(N_r/N_d) \equiv N_b^{sb}$. It's very easy to show that (1) becomes

$$E[D_v] = N_d^{bb} \left[1 - \prod_{i=0}^{N_b^{bb}-1} \left(1 - \frac{s}{N_r - i} \right) \right] + N_d^{sb} \left[1 - \prod_{i=0}^{N_b^{sb}-1} \left(1 - \frac{s}{N_r - i} \right) \right] \quad (3)$$

3 Proof for the skewed (general) distribution

The easiest way to build the formula is to reverse the problem: calculating the probability of *not* selecting balls.

3.1 Probability of not selecting a bucket

This is the probability of selecting no ball from the bucket; using NS as a shorthand for "Not Selecting", let's define:

$$Pnib(k) \equiv P(NS(ball_0) \cap NS(ball_1) \cap ... \cap NS(ball_{N_b(k)-1}))$$

Let's build the combined probability by considering each ball in turn.

Assuming, of course, that we pick balls completely at random, for the first ball we have

$$P(NS(ball_0)) = \frac{N_r - s}{N_r}$$

Thanks to Bayes' Theorem:

$$P(NS(ball_0) \cap NS(ball_1)) = P(NS(ball_0)) * P(NS(ball_1)/NS(ball_0))$$

Since we pick balls *without replacement*, we know that the second ball has an higher probability to be selected, now that we know that the first ball has escaped the hand of the selecter. In fact we have still s "chances to be picked" in a bag that now has one ball less $(N_r - 1)$, so :

$$P(NS(ball_1)/NS(ball_0)) = \frac{(N_r - 1) - s}{(N_r - 1)}$$

thus

$$P(NS(ball_0) \cap NS(ball_1)) = \frac{N_r - s}{N_r} \frac{N_r - 1 - s}{N_r - 1}$$

Iterating until the last ball, we have

$$Pnib(k) = \frac{N_r - s}{N_r} \frac{N_r - 1 - s}{N_r - 1} \frac{N_r - 2 - s}{N_r - 2} ... \frac{N_r - (N_b(k) - 1) - s}{N_r - (N_b(k) - 1)}$$

or

$$\boxed{Pnib(k) = \prod_{i=0}^{N_b(k)-1} \left(\frac{N_r - i - s}{N_r - i} \right) = \prod_{i=0}^{N_b(k)-1} \left(1 - \frac{s}{N_r - i} \right)} \quad (4)$$

3.2 Combining the buckets

If we define $I(k)$ as the *indicator variable* of the event "k-th bucket selected", ie

$$I(k) \equiv \begin{cases} 1 & \text{if the } k\text{-th bucket is selected,} \\ 0 & \text{if the } k\text{-th bucket is not selected.} \end{cases}$$

We have that

$$D_v = \sum_{k=0}^{N_d-1} I(k)$$

and

$$E[D_v] = E\left[\sum_{k=0}^{N_d-1} I(k)\right] = \sum_{k=0}^{N_d-1} E\left[I(k)\right]$$

But

$$E\left[I(k)\right] = 0 * P(I(k)=0) + 1 * P(I(k)=1) = P(I(k)=1) = 1 - Pnib(k)$$

so, using the (4) result, we get

$$E[D_v] = \sum_{k=0}^{N_d-1} (1 - Pnib(k)) = \sum_{k=0}^{N_d-1} \left(1 - \prod_{i=0}^{N_b(k)-1} \left(1 - \frac{s}{N_r - i}\right)\right)$$

which is the (1) general formula.

4 Test with brute-force simulations

The formulae have been tested against the results provided by brute-force simulators, i.e. programs [1] that simulates selecting balls from a bag in every possible way, and compute the average D_v observed.

The match (over thousands of test cases, including "border" ones) has always been exact, minus of course the rounding and truncation errors inherent in IEEE 754 double floating-point arithmetic.

For example, for our example bag [2] :

s	$E[D_v](formula)$	$avg[D_v](bruteforce)$	$abs(difference)$
0	0.0	0.0	0.0
1	1.0000000000000002	1.0	$2.220446049250313E-16$
2	1.772727272727273	1.7727272727272727	$2.220446049250313E-16$
3	2.368181818181818	2.368181818181818	0.0
4	2.826262626262626	2.8262626262626265	$4.440892098500626E-16$
5	3.1780303030303028	3.178030303030303	$4.440892098500626E-16$
6	3.4469696969696972	3.446969696969697	$4.440892098500626E-16$
7	3.6502525252525255	3.650252525252525	$4.440892098500626E-16$
8	3.8000000000000003	3.8	$4.440892098500626E-16$
9	3.9045454545454548	3.9045454545454548	0.0
10	3.9696969696969697	3.9696969696969697	0.0
11	4.0	4.0	0.0
12	4.0	4.0	0.0

Note that the maximum difference has been $4.44 * 10^{-16}$.

4.1 Related Works

After deriving these formulae using elementary statistical considerations, I discovered that a formula for the uniform case has already been published - see "'S.B. Yao, Approximating Block Accesses in Database Organizations"' in "'Communications of the ACM, April 1977, Volume 20, Number 4"'. Since the problem discussed here arises in many applications for sure, I'm quite sure that even the general case has probably already been investigated in depth and published somewhere, and I would appreciate to get the reference/paper if that is the case. My email is alberto.dellera@gmail.com. Thanks!

Written using a LaTeX environment (tools MiKTeX and TeXnicCenter)
see http://www.artofproblemsolving.com/LaTeX/AoPS%5FL%5FDownloads.php

[1] DistinctBalls.java contains the ones I've developed and used, as well as Java versions of the formulae.
[2] Generated by : java DistinctBalls skewed exhaustive_latex 5 2 3 2 12