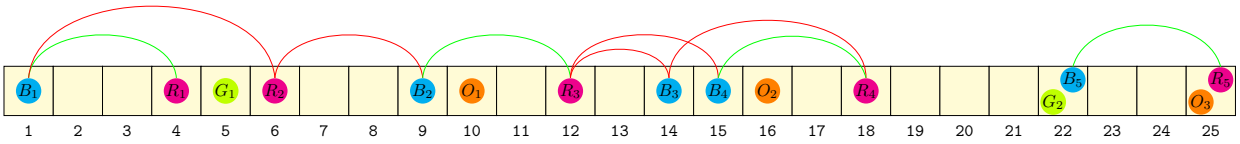
1 The problem statement

I would like to create statistical test that detects enrichment or over-representation of structures such as pairs, triplets, quadruplets, etc. in a given group of obejcts compared to the other group.

The tricky part is to take into account the substructures. For example, I'm interested in over-representation of quadruples of object and I suppose I should somehow take into consideration and enclose information about lower structures like pairs or even the frequencies of individual objects.

But let's start at the very beginning. I have N groups represented by ordinary one-dimensional arrays each of length L_i. I also have colorful balls arranged in a certain way in groups cells. Balls may co-occupy the same cell, but only one ball of a given color per cell (i.e. only balls of different color may co-occupy the same cell). The whole situation is given and we have full information about it (more about the whole arrangement in section A).



Now, the example for pairs. Let's say, I'm particulary interested in pairs of blue and red balls that preseve a fixed structure composed of:

- **an order** blue is on the left hand side and red on the right and
- **a distance** between this two cells containing this balls is set to e.g. 3.

In the picture above, with green arcs there're marked pairs that preserve the structure, whereas red arcs denote pairs violating order or distance requirement.

Subsequently, I choose one group, which will be called a foreground (e.g. $F = G_1$) and the remaining groups all together will form the background (e.g. $B = \bigcup_{2 \le i \le N} G_i$).
Of course, I don't merge background groups side by side and don't accept pair starting in one group and with the other ball in another group.

Now, I'm looking for a rather rudimental statistical test for checking:

- Null hypothesis H_0: the probability of occurrence of structured (maintained order and distance) pair of blue and red balls is the same or lower in the foreground group $F = G_1$ than the background $B = \bigcup_{2 \le i \le N} G_i$ (other groups all together)

  versus

- Alternative H_A: probability of such structured pair is higher in the foregorund.

Obviously, I'm able to create test for enrichment of single objects, but I have internal premonition that I should take into account lower structures like pairs and even the frequencies of individual objects of which these conformations are composed.
In my tests, there's nothing special about pairs. I would like to adapt them (or whichever other test) to handle higher structures e.g. triplets, but I feel really bad to literally replace pairs with triplets. In the next section it will be explained what I mean by *literally replacing pairs with triplets*.

2 Most trivial approach only giving an vague insight

In this section I'm going to present the most trivial approach only to give a vague idea about the problem. Naturally, this method may be inappropriate also given the data statistics A and resulting assumption to rather use Poisson distribution, but don't bother with that.

For a while, let's assume that occurrence of such structured pair anywhere in the foreground is equally likely, with the same unknown probability p_f for each cell. The same applies to background and unknown probability p_b. After that, we get two sequences of independent random variables $F_i \sim Bernoulli(p_f)$ and $B_i \sim Bernoulli(p_b)$, where $F_i = 1$ denotes occurrence of structured pair starting at i-th cell in the foreground and the same for B_i. All groups are long, so from Central Limit Theorem we get that both the mean of F_is and the mean of B_is have Normal distribution:

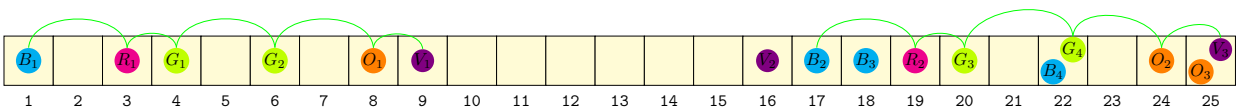
$$\text{(mean of } F_i\text{)} \qquad \overline{F} = \frac{\sum\limits_i F_i}{L_1} \sim Normal(p_f, \frac{p_f \cdot (1-p_f)}{L_1})$$

$$\text{(mean of } B_i\text{)} \qquad \overline{B} = \frac{\sum\limits_i B_i}{\sum\limits_{2 \le i \le N} L_i} \sim Normal(p_b, \frac{p_b \cdot (1-p_b)}{\sum\limits_{2 \le i \le N} L_i})$$

Because I don't know anything about variations, thus I use Welch's T test (but this selection of Welch's T test is not crucial and isn't a problem at all).

However, the main problem lies here: In this test there's nothing special about pairs and that can lead to, in my opinion erroneous, conclusion one can literally replace pairs with any other higher structures (e.g. F_i and B_i will denote start of triple instead of pair and we change nothing else). For me personally, that's a suspicious idea and at least such test needs some kind of adjustment. In the next section 3 I'm going to present what kind of adjustment I'm talking about.

Let's have a look on two bigger structures consisting of six objects. On the left hand side we have refernce structure and on the right the occurence of this structure interlaced with additional objects, but that's OK.

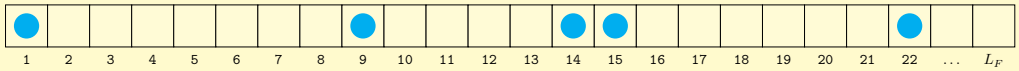


Personally, if I'm looking for over-representation of such larger structures like this hexatruple (sextuple), I think I should consider the incidences of constitutive pairs.
If a particular pair appears exceptionally rarely, but most of the time as a constitutive component of triples, quadruples or even whole structure, that's gives an valuable insight.
On the other hand, if a particular pair occures very frequently, but not as a part of any larger structure (not even the whole hexatruple, but rather component triple or quadruple), I believe I should penalize or diminish occurrences of this pair.

3 Solution that works properly only for pairs

In this section, I'm going to present the solution that **works only for pairs**, but what's most important **incorporates the information about unstructured pairs**.
Nevertheless, at first I have to recall a quite well known property which leads to statistical test for comparing means of two Poisson distributed random variables. Of course I'm aware that there're some more powerful tests, but I use here C-test in a purpose.

Property 1. *Conditional C-test by Przyborowski & Wileński, 1940*
We have two random variables $A \sim Poisson(\lambda_A)$ *and* $B \sim Poisson(\lambda_B)$ *describing number of occurrence of a balls (of single color) in two arrays of lengths* N_A *and* N_B *respectively (like in the picture below):*



Assuming that a *and* b *denote the actual (empirically measured) values of* A *and* B *and let* $n = a + b$*, it's easy to show that the conditional distribution* $A|A + B = n$ *has binomial distribution. Precisely*

$$P(A \geq x|A + B = n) = \sum_{i=x}^{n} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} \quad \textit{where} \quad p = \frac{\lambda_A}{\lambda_A + \lambda_B}$$

Because these arrays are of different lengths $N_A \neq N_B$*, so we're interested in comparing their rates* μ_A *and* μ_B *where* $\lambda_A = \mu_A \cdot N_A$ *and* $\lambda_B = \mu_A \cdot N_B$*, what gives*

$$p = \frac{\mu_A \cdot N_A}{\mu_A \cdot N_A + \mu_B \cdot N_B}$$

Under the null hypothesis of equal rates $\mu_A = \mu_B$ *this probability reduces to*

$$p = \frac{N_A}{N_A + N_B}$$

Now, it's straightforward to test the equality of rates using the upper tail of cumulative distribution function of binomial distribution, rejecting H_0 *if* $p\text{-value} = P(A \geq a|A + B = n) \leq \alpha$ *(*α *is a significance level e.g. 0.05)*

Proof 1. *According to joint probability distribution we can state*

$$P(A=a, B=b) = \frac{\lambda_A^a \cdot e^{-\lambda_A}}{a!} \cdot \frac{\lambda_B^b \cdot e^{-\lambda_B}}{b!}$$

Assuming that $n = a + b$, $\mu = \lambda_A + \lambda_B$, $p = \frac{\lambda_A}{\lambda_A + \lambda_B}$ *it's easy to rewrite it as*

$$P(A=a, B=b) = [\frac{\mu^n \cdot e^{-\mu}}{n!}] \cdot [\frac{n!}{a! \cdot (n-a)!} \cdot p^a \cdot (1-p)^{n-a}]$$

The first factor is responsible for reaching the total sum of n*, whereas the other splits* n *into* a *and* b*. Now it's trivial to see*

$$P(A=a, B=b | A+B=n) = \frac{n!}{a! \cdot (n-a)!} \cdot p^a \cdot (1-p)^{n-a}$$

Finnally gathering $\binom{n}{a} = \frac{n!}{a! \cdot (n-a)!}$ *and summing up proves that*

$$P(A \geq x | A+B=n) = \sum_{i=x}^{n} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i}$$

Now I must introduce some designations (frankly speaking I've changed the variable names from an article by *Jankowski et al., 2013* , but don't bother with this article).

- $R_{MAX} = 100$ maximal radius or distance within which two balls can be referred as a pair. As I mentioned before, I'm going to use unstructured pairs violating the distance requirement, but of course in a reasonable range, e.g. an one ball from the outset of a group and a second one from the very end **cannot or should not** be referred as a pair.
- F_S - the total number of pairs of blue and red preserving the structure (given order and distance = 3) in the foreground.
- B_S - the total number of pairs of blue and red preserving the structure (given order and distance = 3) in the background.
- F_U - the total number of pairs of blue and red in the foreground, but **without** the requirement to maintain the structure, e.g. in any order and at any distance $\leq R_{MAX}$.
- B_U - the total number of pairs of blue and red in the background, but **without** the requirement to maintain the structure, e.g. in any order and at any distance $\leq R_{MAX}$.
- MAX_{FS} - the maximal theoretical number of pairs of blue and red retaining the structure that can entirely fit into foregroud. **This value is not based on balls arrangement, but calculated from the group length and the structure length.**
  In our example that's $MAX_{FS} = L_1 - 4 + 1$ (group length - structure length + 1).
- MAX_{BS} - the maximal theoretical number of pairs of blue and red retaining the structure that can entirely fit into backgroud. In our example that's $MAX_{BS} = L_B - 4 + 1$. Here, only for the simplicity L_B denotes the background length, however it will be explained in a moment.

- MAX_{FU} - the maximal theoretical number of pairs of blue and red in the foreground **without** the requirement to maintain the structure, but still with the requisite for distance $\leq R_{MAX}$. **As in case of MAX_{FS} and MAX_{BS}, this value is also artificially computed and not read from balls arrangement.** In our example that's $MAX_{FU} = L_1 \cdot (2 \cdot R_{MAX} + 1) - R_{MAX} \cdot (R_{MAX} + 1)$. This formula works properly as long as $R_{MAX} \leq L_1$, but this is easily satisfied.
  It behove me to explain a derivation of this formula. Let's say, I trying to exhaust all possible configurations of unstructured pairs of blue and red by choosing at first position for blue ball and then cell for red one with respect thereto. I can place blue ball in L_1 positions. If blue one is in the middle of the group, the place for red can be picked out in $2 \cdot R_{MAX} + 1$ different ways (in R_{MAX} cells on the left + in the same call + in R_{MAX} cells on the right). One cannot forget to deduct these arrangements that fall off when approaching the margins. By placing blue at R_{MAX}-th cell we're losing 1 possibility, ..., by placing blue at 1st cell we're losing up to R_{MAX} possibilities for red one, which gives in total $\frac{R_{MAX} \cdot (R_{MAX}+1)}{2}$ to deduct only at one side, so we must double that.

- MAX_{BU} - the maximal theoretical number of pairs of blue and red in the background **without** the requirement to maintain the structure.
  As previously $MAX_{BU} = L_B \cdot (2 \cdot R_{MAX} + 1) - R_{MAX} \cdot (R_{MAX} + 1)$.
  Here, I have to admit to unstated premise that the background consist of one and only group of length L_B. At the beginning, in section 1 I granted that the foreground is $F = G_1$ and the background consist of all other groups all together $B = \bigcup_{i=2}^{N} G_i$ without merging them side by side. To be very precisely, I should rather state
  $MAX_{BU} = (\sum_{i=2}^{N} L_i) \cdot (2 \cdot R_{MAX} + 1) - (N-1) \cdot R_{MAX} \cdot (R_{MAX} + 1)$ only if $\forall_i R_{MAX} \leq L_i$.
  This very scrupulous approach brings completly nothing to the solution explanation except for the unnecessary notational burden. That's why I'm using simply L_B.

3.1 Straightforward application of C-test (warm-up only)

According to conditional C-test by *Przyborowski & Wileński* under the null hypothesis of equal rates, one can compute the *p-value* - the probability of observing **at least** F_S structured pairs in the foreground as follows:

$$\textbf{p-value} = PBINOM(F_S - 1,\ size = F_S + B_S,\ prob = p,\ lower.tail = FALSE)$$

where probability of success in a single trial is equal to

$$p = p_1 = \frac{MAX_{FS}}{MAX_{FS} + MAX_{BS}}$$

In most statistical environments (like R) PBINOM (CDF for Binomial distribution) computes upper tail in a strict manner $P(X > x)$ while I need a non strict version $P(X \geq x)$, thus I use $F_S - 1$ for a technical sake only.
If p-value $\leq \alpha = 0.05$, the null hypothesis is rejected.

3.2 Supreme adjustment taking into account unstructured pairs

Ultimately, we move on to the most imporatnt conclusion, that the densities of occurrences of unstructured pairs between the foreground and the background may significantly vary, so we can compensate this by simply changing single trial probability of success Why? Because how can I expect many structured pairs in a particular group if this group does not have many pairs without the restriction on the mutual order and distance.

Assuming the following empirical probabilities for blue and red pairs

- $Prob_{FU} = \frac{F_U}{MAX_{FU}}$ - pobability of occurrence of unstructured pair in the foreground.
- $Prob_{BU} = \frac{B_U}{MAX_{BU}}$ - pobability of occurrence of unstructured pair in the background.

one can adjust the probability of success in single trial:

$$p = p_2 = (\frac{MAX_{FS}}{MAX_{FS} + MAX_{BS}}) \cdot (\frac{Prob_{FU}}{Prob_{BU}})$$

The second factor $\frac{Prob_{FU}}{Prob_{BU}}$ compensate the difference in densities of unstructured pairs between the foreground and the background.

I think that, this weighted probability of success in single trial utterly shows the validity of incorporating unstructured pairs.

Moreover, in my opinion one should find similar legitymacy in enclosing information about constitutive pairs in solution for bigger structures.

Question 1/2
Do you think such adjustment for unstructured pairs is legitimate and conducted in proper manner?

What's about such adjustment in tests for enrichment of bigger structures? Do I need to filter out also instances which **do not preserve strictly requirements** and similarly incorporate such information?

For me personally, incorporating unstructured pairs acts more or less as incorporating information about constitiuent substructures,

4 The Ultimate Question

Question 2/2
How to create a statistical test for enrichment of bigger structures like triplets, quadruples, etc.?
Do you agree that I need to somehow enclose information about constitutive pairs or even individual objects? If yes, show me how should it be done.
Or maybe on the contrary, I should simply threat bigger structures as a whole and count occurences only of full instances? In this approach, can one make an assumption that every test for enrichment of single objects could be used by literally replacing objects with higher structures?
I'm looking not only for complete test for over-representation, but at least a guidance what kind of information to involve and how to balance or weigh leverage of various components. With such knowledge I will be able to create the final test for enrichment by my own. I can always adjust general Wald test and combine it with a some well-known test for enrichment.

A Data statistics

Group lengths	
Total number of groups	44
Average group length (number of cells)	1 954 279
Standard deviation of group length	1 571 307
Minimal group length	163 480
Maximal group length	7 298 852
Total sum of lengths of all groups	85 988 316

In the next table I present statistics of **single ball** occurrences. The quantities of appearanace of **pairs** are definitely much smaller, thus assuming Poisson distribution is acceptable and even desirable. If we focus only on pairs of fixed order and spacing we get a few thousands or at most several dozens of such structured couples through 86 millions of cells in total, so they're rather rare events. The bigger the structure the lesser occurrences we get.

Balls statistics (occurrences are summed up across all groups)	
Max. no. of different ball colors in structure	7 *
Average single color occurrences	2 835 499
Standard deviation of single color occurrences	3 646 186
Minimal single color occurrences	31 194
Maximal single color occurrences	12 177 072

* This number 7 denotes the maximal number of different colors of which the reference structure can be build. Generally, there're dozens of different colors, but if I'm taking up for examination of one particular reference structure, I narrows this amout to max. seven different colors and let's say max. 9 balls in total (simply for some colors there're repetitions, recall the hexatruple and two green balls).

A.1 How balls are spread in cells

Maybe in the final solution we would have to make use of frequencies of even individual balls, so here I present how they're distributed and maybe would be able to infer their probability in each cell separately.

As I mentioned before, the whole layout of balls in cells is given to us and we have complete knowledge about it, but we have available only one single dataset. There's no pure randomness in this positioning process in the sence that I cannot gather another set of data next time. The balls will apear at the same position according to a startegy below:

First of all, I choose one color that apears in the reference structure and in a certain manner I choose a *threshold* $\in [0, 1]$.
Then, I go through all groups and for each cell I compute the specific *Score* $\in [0, 1]$ which is compared with this threshold. If and only if $Score \geq threshold$ I place ball of this particular color. This score gives an intrinsic affinity of balls of that color for given cell and it's in a range [0, 1], so maybe it can be used as a some kind of measure. For sure it's a measure because it meets all measure axioms. The only doubt is whether we can use it as a probability measure, but for sure it's a measure in more general sense. I do this procedure for each color from reference structure individually and the threshold is choosen in each case.

Below there's an example of blue balls placement. The first picture is only a repetition of our first example, whereas in the second chart we can see scores for individual cells and threshold set to 0.8. The score value is reached for cell: 1, 9, 14, 15, and 22, thus blue balls are placed in them.