Matching Initials: A Surprising Pattern in Spouse Names

Gaetan Ecrepont

May 2024

Abstract

The selection of a life partner is influenced by a complex interplay of sociological, psychological, and anthropological factors, ranging from emotional connection and social norms to cultural traditions. This study introduces a novel factor to the discourse: the matching of initials between spouses. Utilizing a comprehensive dataset from the New York City Marriage Index, covering marriages from 1950 to 2017, we analyze the prevalence and statistical significance of this phenomenon. Our findings suggest a notable preference for partners with matching initials, revealing a subtle yet statistically significant pattern.

1 Introduction

This study investigates whether the initials of spouses' names might influence marital decisions, a topic not widely considered in previous research. Using a comprehensive dataset from the New York City Marriage Index, which includes millions of marriage records spanning from 1950 to 2017, we analyze the occurrence of matching initials among spouses.

Our methodology compares the observed frequencies of matching initials to those expected under random pairing. Initial results show a higher occurrence of matching initials than expected by chance, indicating a potential preference for this trait. This paper focuses on quantitatively assessing these observations, considering the impact of various biases such as endogamy which might affect the independence of spouse name selection.

The goal is to present a clear statistical analysis of whether matching initials are a factor in spouse selection, without making broad generalizations about its significance or implications.

To be clear, we are not claiming that matching initials are an important factor in marriage outcomes. However, we argue that it does have a small impact which, despite being much weaker than other well-known decision factors, is large enough for us to demonstrate its significance.

This paper will be organized in two parts. We will first present the data and some general statistics on it, then we will deploy statistical tests to acertain the

significance of our findings.

2 Data

2.1 Source

The dataset we'll use comes from `nycmarriageindex.com`. It is a public record of almost all the marriages registered in New York City between 1950 and 2017. There are roughly 4.6M rows with 6 columns: bride's first and last name, groom's first and last name, license city, and license year.

2.2 Cleaning

We have cleaned the data from about 4.6M records to about 3.6M. Most notably, we filtered out the following rows:

- rows containing missing values
- rows containing a first name (groom and/or bride) with less than 5 occurrences in the dataset (such names are considered typos or exceedingly rare)
- rows with first names consisting of 3 characters of less (such names are almost always acronyms)
- rows containing a first name (groom and/or bride) which includes characters outside the alphabet (e.g. a number or a hyphen)

2.3 Flaws

Despite our cleaning, these dataset still suffers from a few flaws:

1. biased towards NYC inhabitants
2. some names are mispelled
3. a handful of years in our dataset seem to be missing some records as they have a lot less marriages than the rest

Although points 2 and 3 have virtually no consequence on our findings, point 1 is important and must be kept in mind. Indeed, NYC has a notoriously multicultural population and as we shall see, endogamy creates certain correlations in the names (and hence the initials) of the bride and groom.

3 Statistics

3.1 Notations

Let's introduce some notations which we will hold to throughout this paper:

- $\mathcal{A} = \{A, B, C, \ldots, Z\}$ is the latin alphabet
- Ω_G (resp. Ω_B) is the set of all grooms (resp. all brides)
- $G : \Omega_G \to \mathcal{A}$ (resp. $B : \Omega_B \to \mathcal{A}$) is the random variable which to a given groom (resp. bride) associates its first name initial. G (resp. B) thus follows a categorical distribution over $\mathcal{A}$ which we will denote $\mathbb{P}_G$ (resp. $\mathbb{P}_B$)
- the joint distribution $(G, B) : \Omega_G \times \Omega_B \to \mathcal{A} \times \mathcal{A}$ will be denoted $\mathbb{P}_{GB}$ such that our dataset $\mathcal{D} = \{(G_i, B_i)\}_{1 \leq i \leq N}$ of N mutually independent (G, B) observations can be seen as a sample from $\mathbb{P}_{GB}^{\otimes N}$, while $\mathcal{D}_G = \{G_i\}_{1 \leq i \leq N}$ (resp. $\mathcal{D}_B = \{B_i\}_{1 \leq i \leq N}$) can be seen as a sample from $\mathbb{P}_G^{\otimes N}$ (resp. $\mathbb{P}_B^{\otimes N}$)
- $\forall (\alpha, \beta) \in \mathcal{A} \times \mathcal{A}, p_\alpha^G = \mathbb{P}_G(\alpha)$, $p_\beta^B = \mathbb{P}_B(\beta)$, and $p_{\alpha\beta} = \mathbb{P}_{GB}(\alpha, \beta)$. In particular, $p_{\alpha\beta} \not\equiv p_\alpha p_\beta$ *a priori*.

3.2 Preliminary warning

The large size of our dataset ($N \sim 3.6 \cdot 10^6$ records) allows us to expect asymptotic theorems to be valid for our tests. However, extra caution must be used as we will be dealing with low-probability events, especially for rare letter couples. Indeed, although individual joint probabilities should theoretically be of order of magnitude $p_{\alpha\beta} \sim \frac{1}{26} \times \frac{1}{26} \sim 10^{-3}$, in practice some initials are much rarer than other. For instance, the couple of initials (U,Q) only contains one instance in our dataset! Likewise for the couple (Q,U). For such couples, we cannot expect to do much statistics as they are not even statistically significant. Indeed, for a given couple $(\alpha, \beta) \in \mathcal{A}^2$, testing for the null hypothesis $p_{\alpha\beta} = 0$ with alternative $p_{\alpha\beta} > 0$, yields the z-score:

$$z = \frac{p_{\alpha\beta}}{\sqrt{\frac{p_{\alpha\beta}(1-p_{\alpha\beta})}{N}}} \underset{p_{\alpha\beta} \ll 1}{\simeq} \sqrt{N p_{\alpha\beta}} = \sqrt{N_{\alpha\beta}}$$

where $N_{\alpha\beta}$ is the number of occurrences of the couple (α, β) in the dataset. Therefore, if we set the level of significance at 3 standard deviations, we need at least $3^3 = 9$ observations of a given couple (α, β) to even consider it. Thus, rare couples such as (Q,U) and (U,Q) can be discarded already. Luckily, we are focusing on matching initials only i.e. couples of the form (α, α) and such couples all have more than 9 instances in the data. In fact, they all have more than 100 observations except for (Q,Q) and (U,U), which have 36 and 14 instances respectively.

Thus, since we are dealing with rare events, we shall use extra caution when using asymptotic theorems.

3.3 First look

Before performing any tests, the most natural way to gauge people's appetence for partners with matching initials is to compare for each initial $\alpha \in \mathcal{A}$ the observed frequency $p_{\alpha\alpha}$ to the expected frequency $p_\alpha^G p_\alpha^B$. If our assumption that people like matching initials is correct, then we should find that $p_{\alpha\alpha} > p_\alpha^G p_\alpha^B$ for most if not all α.

Let's first look at the distribution of G and B, as illustrated in Figure 2. Nothing too surprising here either, J for JOHN and M for MARY/MARIE/MARIA dominate significantly. Initials Q, W, X, Y, Z are rare and account mainly for the Chinese population in NYC. As expected, U is almost non-existent for both genders.

Let's now look at the joint distribution and compare it the observed distribution $\mathbb{P}_{GB}$ to the expected distribution $\mathbb{P}_G \times \mathbb{P}_B$. As mentioned above, we only look at matching initials since 1) they're the relevant object of this study and 2) some couples are so rare that they aren't even statistically significant. The histogram of observed and expected frequencies is given in Figure 1. The absolute and relative difference for each initials is given in Figure 3.

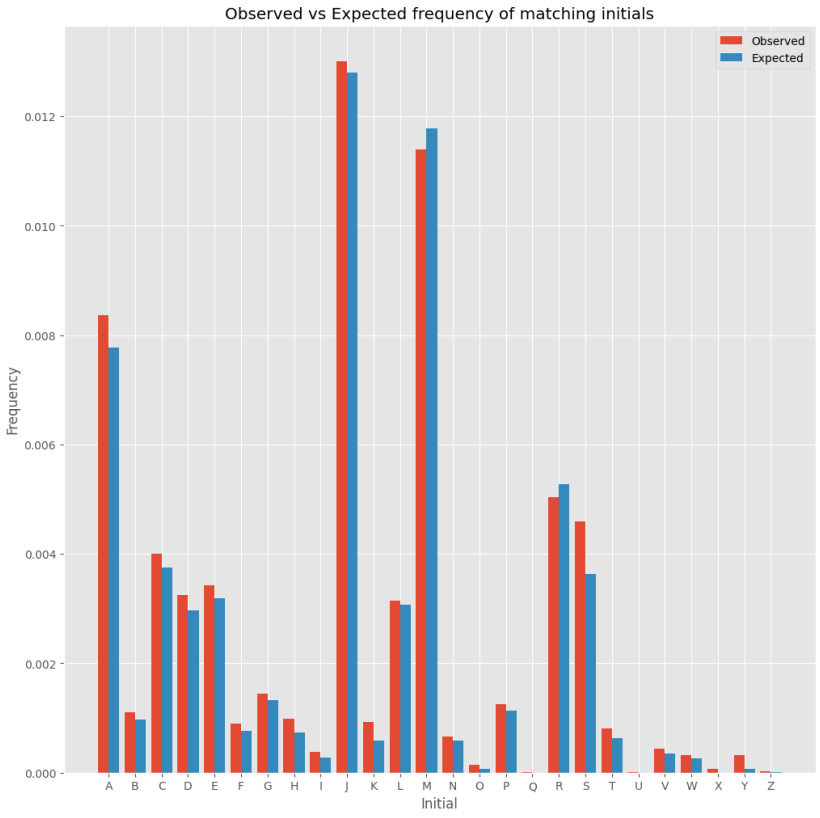

Figure 1: Histogram of the observed and expected frequencies for each couple of matching initials.

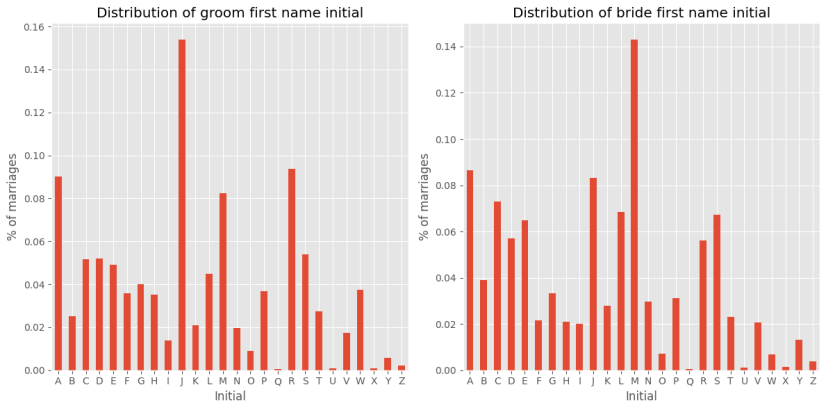

Figure 2: Distribution of G and B, respectively.

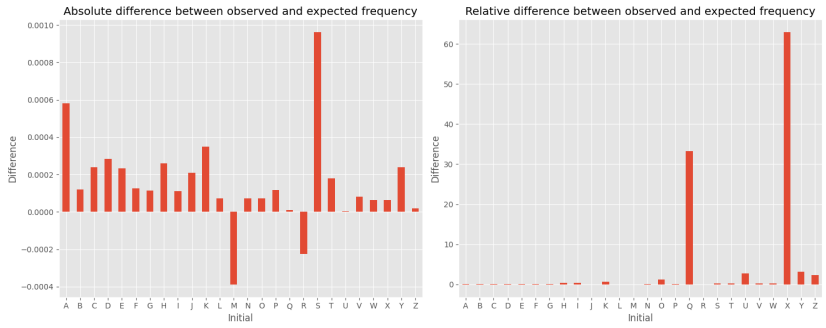

Figure 3: Absolute and relative difference between observed and expected frequencies for each couple of matching initials.

As hoped, observed frequencies are always greater than expected frequencies, except for initials M and R.

We will now use statistical theory to assess the significance of the observed discrepancies.

3.4 Significance of the deviations found

Before unrolling more powerful tests, we can already give our present findings some intuitive significance. Indeed, under the null hypothesis that people are indifferent to their partner's first name initial, i.e. $H_0 : \forall \alpha \in \mathcal{A}, p_{\alpha\alpha} = p_\alpha^G p_\alpha^B$, then we expect that $S_\alpha = \text{sign}\big(\widehat{p_{\alpha\alpha}} - \widehat{p_\alpha^G}\widehat{p_\alpha^B}\big) \underset{\mathcal{L}}{\simeq} B(\frac{1}{2})$, such that $S = \sum_{\alpha\in\mathcal{A}} S_\alpha \underset{\mathcal{L}}{\simeq} B(26, \frac{1}{2})$. Thus, $p = \mathbb{P}(S \leq 24) \simeq 5 \cdot 10^{-6}$ is the corresponding p-value, which overwhelmingly rejects H_0. However this is a heuristic and we shall now introduce more serious statistical tests.

For a given initial $\alpha \in \mathcal{A}$, we want to test the null hypothesis $H_0^\alpha : p_{\alpha\alpha} = p_\alpha^G p_\alpha^B$ against the alternative $H_1 : p_{\alpha\alpha} \neq p_\alpha^G p_\alpha^B$. Note that we are using two-tailed tests since we saw in the previous part that letters M and R exhibit the opposite of the expected pattern.

Let $\alpha \in \mathcal{A}$ an initial. By definition, the probability of observing a couple (G_i, B_i) with initials α is given by $p_{\alpha\alpha}$. The most natural estimator for this probability is $\widehat{p_{\alpha\alpha}} = \frac{1}{N}\sum_{i=1}^N 1_{G_i=B_i=\alpha}$. Then, by virtue of the CLT:

$$\frac{\widehat{p_{\alpha\alpha}} - p_{\alpha\alpha}}{\sqrt{\frac{p_{\alpha\alpha}(1-p_{\alpha\alpha})}{N}}} \underset{N}{\xrightarrow{\mathcal{L}}} N(0,1)$$

Plugging in the null hypothesis $p_{\alpha\alpha} = p_\alpha^G p_\alpha^B$, we have that

$$\frac{\widehat{p_{\alpha\alpha}} - p_\alpha^G p_\alpha^B}{\sqrt{\frac{p_\alpha^G p_\alpha^B (1-p_\alpha^G p_\alpha^B)}{N}}} \underset{N}{\xrightarrow{\mathcal{L}}} N(0,1)$$

However, we do not know $p_\alpha^G p_\alpha^B$ and therefore cannot use this z-test. One could be tempted to invoke Slutsky's theorem and replace $p_\alpha^G p_\alpha^B$ with its canonical estimator $\widehat{p_\alpha^G p_\alpha^B} = \widehat{p_\alpha^G}\widehat{p_\alpha^B}$ where $\widehat{p_\alpha^G} = \frac{1}{N}\sum_{i=1}^N 1_{G_i=\alpha}$ and $\widehat{p_\alpha^B} = \frac{1}{N}\sum_{i=1}^N 1_{B_i=\alpha}$. However, although we have a large number N of samples, we do not have enough for Slutsky's implicit assumption $N \simeq +\infty$ to hold (cf. subsection 3.2).

We must therefore adopt a more complex approach in which we detail the asymptotics of the estimator $\widehat{p_\alpha^G p_\alpha^B} = \widehat{p_\alpha^G}\widehat{p_\alpha^B}$.

For $1 \leq i \leq N$, let's consider

$$X_i = \begin{pmatrix} 1_{G_i=\alpha} \\ 1_{B_i=\alpha} \\ 1_{G_i=Bi=\alpha} \end{pmatrix}$$

and

$$\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i = \begin{pmatrix} \widehat{p_\alpha^G} \\ \widehat{p_\alpha^B} \\ \widehat{p_{\alpha\alpha}} \end{pmatrix}$$

.

By virtue of the three-dimensional CLT:

$$\sqrt{N}(\bar{X} - \mu) \xrightarrow[N]{\mathcal{L}} N(0, \Sigma)$$

where $\mu = \mathbb{E}[X_i] = \begin{pmatrix} p_\alpha^G \\ p_\alpha^B \\ p_{\alpha\alpha} \end{pmatrix}$ and $\Sigma = \mathrm{Cov}(X_i) = \begin{pmatrix} p_\alpha^G(1-p_\alpha^G) & p_{\alpha\alpha} - p_\alpha^G p_\alpha^B & p_{\alpha\alpha}(1-p_\alpha^G) \\ p_{\alpha\alpha} - p_\alpha^G p_\alpha^B & p_\alpha^B(1-p_\alpha^B) & p_{\alpha\alpha}(1-p_\alpha^B) \\ p_{\alpha\alpha}(1-p_\alpha^G) & p_{\alpha\alpha}(1-p_\alpha^B) & p_{\alpha\alpha}(1-p_{\alpha\alpha}) \end{pmatrix}$

Then, using the three-dimensional delta method with $g : (x, y, z) \mapsto z - xy$, we find that

$$\sqrt{N}\big((\widehat{p_{\alpha\alpha}} - \widehat{p_\alpha^G}\widehat{p_\alpha^B}) - (p_{\alpha\alpha} - p_\alpha^G p_\alpha^B)\big) \xrightarrow[N]{\mathcal{L}} N\left(0, \begin{pmatrix} -p_\alpha^B & -p_\alpha^G & 1 \end{pmatrix} \Sigma \begin{pmatrix} -p_\alpha^B \\ -p_\alpha^G \\ 1 \end{pmatrix}\right) \underset{\text{notation}}{=} N(0, s_\alpha^2)$$

Finally, plugging in the null hypothesis $p_{\alpha\alpha} = p_\alpha^G p_\alpha^B$ and using the canonical estimator for s_α^2 (i.e. we take the formula for s_α^2 and replace each term $p_\alpha^G, p_\alpha^B, p_{\alpha\alpha}$ with its respective canonical estimator), we finally obtain the statistic

$$z_\alpha = \frac{\widehat{p_{\alpha\alpha}} - \widehat{p_\alpha^G}\widehat{p_\alpha^B}}{\sqrt{\frac{\widehat{s_\alpha^2}}{N}}}$$

which follows a standard normal distribution as $N \to +\infty$.

Below are the results for each $\alpha \in \mathcal{A}$.

	$p_{\alpha\alpha}$	$p_\alpha^G p_\alpha^B$	z_α	p-value
A	8.36e-03	7.78e-03	1.33e+01	0.00e+00
B	1.10e-03	9.79e-04	7.15e+00	8.83e-13
C	4.00e-03	3.76e-03	7.68e+00	1.64e-14
D	3.25e-03	2.97e-03	1.01e+01	0.00e+00
E	3.42e-03	3.19e-03	8.04e+00	8.88e-16
F	8.97e-04	7.70e-04	8.32e+00	0.00e+00
G	1.45e-03	1.33e-03	5.98e+00	2.22e-09
H	9.88e-04	7.30e-04	1.62e+01	0.00e+00
I	3.87e-04	2.77e-04	1.09e+01	0.00e+00
J	1.30e-02	1.28e-02	3.96e+00	7.53e-05
K	9.29e-04	5.81e-04	2.25e+01	0.00e+00
L	3.14e-03	3.07e-03	2.65e+00	8.10e-03
M	1.14e-02	1.18e-02	-7.80e+00	6.22e-15
N	6.61e-04	5.88e-04	5.59e+00	2.28e-08
O	1.37e-04	6.29e-05	1.22e+01	0.00e+00
P	1.26e-03	1.14e-03	6.53e+00	6.75e-11
Q	9.90e-06	2.89e-07	5.83e+00	5.51e-09
R	5.04e-03	5.27e-03	-6.51e+00	7.70e-11
S	4.59e-03	3.63e-03	2.92e+01	0.00e+00
T	8.09e-04	6.28e-04	1.25e+01	0.00e+00
U	3.85e-06	1.01e-06	2.76e+00	5.75e-03
V	4.40e-04	3.57e-04	7.69e+00	1.47e-14
W	3.21e-04	2.59e-04	6.79e+00	1.13e-11
X	6.30e-05	9.86e-07	1.49e+01	0.00e+00
Y	3.16e-04	7.61e-05	2.62e+01	0.00e+00
Z	2.78e-05	8.45e-06	7.03e+00	2.06e-12

As hoped, we obtain statistically significant results for all letters, though M and R gives results opposite to what we expected. These two exceptions are explained below.

3.5 Explaining the outliers M and R

Let's first focus on M only to understand what is going on. The negative z-score can seem puzzling at first, but after carefully looking at the actual first names starting with M (for both groom and bride) and focusing on the most common ones, we realize that some of them are English and others are foreign, and that names' origins have a large impact on marriage outcomes. Indeed, it appears that people having catholic English names are more likely to marry one another, just as people having Spanish/Portuguese/Italian names. This is endogamy at work.

Interestingly, we find that the most common groom name starting with M is MICHAEL, and the most common bride name starting with M is MARIA.

Clearly, a MARIA is overwhelmingly likely to be of Italian descent, while a MICHAEL probably isn't (Michael Corleone excepted). What if endogamy in NYC was stronger than one might imagine and Italians strongly prefer marrying Italians and likewise for the other ethnic groups? A simple way to check this is to reason on names instead of initials. But there are far too many names to look at. Instead, let's focus on the 10 most common groom and bride names. We can again create a contingency table and then use it to compare the expected joint distribution (under the assumption of independence) to the observed joint distribution. More precisely, we're doing that same chi-square test as in the first test but this time looking at the data subset $\mathcal{D}_{10} = \{(g,b) \in \mathcal{D}_{\text{names}} | (g,b) \in G_{10} \times B_{10}\}$ where G_{10} (resp. B_{10}) is the set of the 10 most common groom (resp. bride) names and $\mathcal{D}_{\text{names}}$ is our initial dataset of first names.

The null hypothesis is that the groom name G and the bride name B are independent, and we perform a chi-square test for independence, we obtain a p-value of less than 10^{-6}, indicating that people dare care about their spouse's name (and not just initial).

Let's now look for each pair $(g,b) \in G_{10} \times B_{10}$ at the relative difference between the observed frequency and the expected frequency, i.e. $\Delta_{gb} = \frac{p_{gb} - p_g p_b}{p_g p_b}$ where p_g (resp. p_b) is the observed frequency of the groom name g (resp. bride name b) and p_{gb} is the observed frequency of (g,b) marriages.

If our endogamy hypothesis is correct, we expect to find positive Δ_{gb} values for names of the same origin, and conversely we expect to find low Δ_{gb} values for names of different origins. The heatmap of Δ_{gb} is given in Figure 4.

Figure 4: Relative deviation from expected joint distribution for the top 10 groom and bride names in the dataset. ** indicates 5% confidence level and *** indicates 1% confidence level.

The results are quite eloquent. The first two cells that stand out are the couples (JOSE,CARMEN) and (JOSE,MARIA), which are a lot more likely than expected (x4.73 and x3.29, respectively). This corresponds to New Yorkers with Spanish/Portuguese/Italian origins marrying each other. In addition, brides named MARIA or CARMEN are less likely than expected to marry any other of the top 10 groom names (except DAVID). Likewise, a groom named JOSE is a lot less likely than expected to marry any other of the top 10 bride names.

Similarly, grooms and brides with catholic English names are likely to marry each other. Interestingly, certain couple seem more probable than others, such as (JOSEPH,MARY) or (DAVID,SUSAN).

With the strong impact of endogamy in mind, we can now understand the unexpected z-scores obtained for initials M and R.

- for M, there are many MARY and MARIE which will overwhelmingly marry JOHN, JOSEPH, ROBERT, and other names not starting in M (with the exception of MICHAEL) ; there are also many MARIA which will marry JOSE and other names not starting in M - for R, there are many ROBERT

(second most represented groom name) which will mary MARY, MARIE, and other names not starting in R (the most popular bride name starting with R is ROSE and it comes at the 21st spot, way below MARY and MARIE for instance)

Thus, the endogamy effect overrides the preference for matching initials for names starting with M and R.

One may wonder why this happen for M and R only. In this case of M, this is probably because the names MARY and MARIA have a marked religious connotation and we can thus expect that the endogamy effect will have a much stronger impact on women carrying such names. ROBERT is less religiously connoted but one can guess that the same applies.

Note that using this reasoning, we might have expected a negative z-score for J matching initials because of the groom name JOHN, which usually goes with MARY or MARIE and not a bride name starting in J. In fact, if we look again at the z-score plots, we see that although positive, the J z-score is among the lowest, proving that the endogamy does play a strong role in this case, although it does not override the matching initials preference.

3.6 Controlling our method

The results we have obtained are quite strong, but maybe too much so. What if there was something biased about our method, which we didn't see? For instance, it is not completely clear how the endogamy effect impacts our results.

One way to control our results is to repeat the same experiment but take other letters than the initials. For instance, we can look at couples (G'_i, B_i) where this time G'_i is the last letter of the groom's first name. In this case, we do not expect a preference for matching letters. Below is the table we obtain if we look at the z_α scores for each couple of matching letters.

	$p_{\alpha\alpha}$	$p_\alpha^G p_\alpha^B$	z_α	p-value
A	2.80e-03	2.40e-03	1.55e+01	0.00e+00
B	9.11e-05	8.05e-05	2.16e+00	3.05e-02
C	4.86e-04	4.92e-04	-5.37e-01	5.91e-01
D	5.86e-03	5.43e-03	1.16e+01	0.00e+00
E	6.19e-03	6.20e-03	-2.29e-01	8.19e-01
F	5.94e-05	5.05e-05	2.23e+00	2.57e-02
G	1.56e-04	2.63e-04	-1.64e+01	0.00e+00
H	9.03e-04	8.64e-04	2.50e+00	1.24e-02
I	1.89e-04	1.75e-04	1.97e+00	4.92e-02
J	4.04e-05	5.45e-05	-4.32e+00	1.55e-05
K	9.64e-04	8.25e-04	8.83e+00	0.00e+00
L	6.60e-03	6.43e-03	4.45e+00	8.48e-06
M	4.27e-03	4.44e-03	-5.46e+00	4.78e-08
N	4.76e-03	4.67e-03	2.64e+00	8.23e-03
O	9.50e-04	6.36e-04	2.08e+01	0.00e+00
P	1.92e-04	1.53e-04	5.42e+00	5.87e-08
Q	0.00e+00	9.99e-08	-2.20e+01	0.00e+00
R	3.24e-03	3.34e-03	-3.47e+00	5.30e-04
S	6.02e-03	6.97e-03	-2.52e+01	0.00e+00
T	1.15e-03	1.19e-03	-2.08e+00	3.75e-02
U	3.30e-06	1.97e-06	1.40e+00	1.61e-01
V	3.77e-05	3.20e-05	1.80e+00	7.11e-02
W	6.44e-05	5.74e-05	1.68e+00	9.30e-02
X	5.50e-06	6.91e-06	-1.15e+00	2.50e-01
Y	9.00e-04	9.12e-04	-8.27e-01	4.08e-01
Z	2.61e-05	1.12e-05	5.59e+00	2.32e-08

Let's focus on the z_α scores only and plot them on Figure 5.

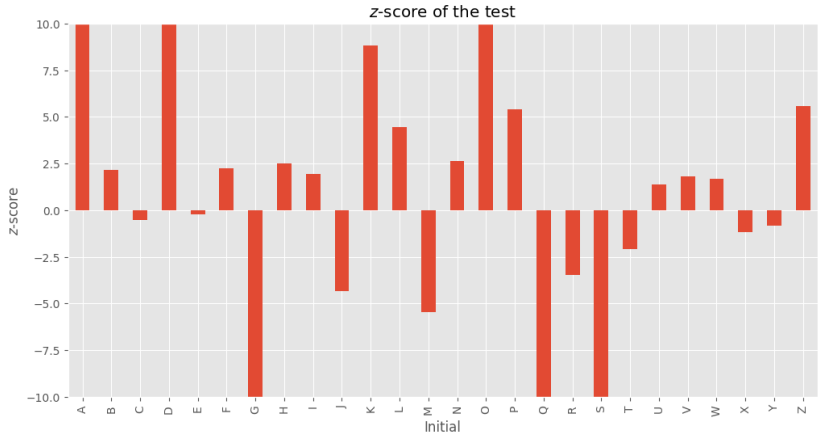

Figure 5: z_α for $\alpha \in \mathcal{A}$, this time looking at couples $(G'i, B_i)$ instead of (G_i, B_i).

We see that the results are a lot more evenly distributed i.e. we have matching couples that appear more than expected, and we also have matching couples that appear less than expected. However, we find high z-scores for many letters, which is surprising. The most plausible hypothesis is that there are still endogamy effects which we have not controlled for. One possible extension to this study would be to remove this presumed endogamy name bias.

4 Conclusion

This study utilized public marriage records from New York City to explore the occurrence of matching initials among spouses. The analysis confirmed a measurable, statistically significant preference for matching initials in most cases, except for the initials M and R. The deviation observed with these letters suggests the influence of endogamy, highlighting the complexity of factors that can affect spouse selection based solely on name initials.

It's important to keep in mind that although the large sample size of our data allows for a certain conclusiveness, this data concerns NYC only and consequently it is not representative of the contemporary world as a whole. It would be interesting to look at other marriage records, in particular in places where names hold no or little cultural significance (if such places exist...) For such places, the name-endogamy effect should vanish and this time we should see that people prefer matching initials for all $\alpha \in \mathcal{A}$. Likewise, it would be interesting to access marriage records of earlier time periods when endogamy was arguably stronger and therefore we might observe a certain vanishing of the matching initials effect as a consequence.