Dude, Where's My Key?

A Bayesian Approach to the Determination of Mythic+ Keystone Distribution in World of Warcraft

William McWhorter

December 8, 2022

Abstract

In the MMORPG World of Warcraft, players can run Mythic+ keystones. The keystones come from a pool of dungeons that varies by season, but is currently at a size of 8. To determine the distribution of the categorical probabilities that keystones are assigned dungeons, a multinomial likelihood was combined with Dirichlet conjugate prior to form a posterior distribution. This distribution was then sampled and used to create posterior credible intervals conditional on keystone level, affix, or affix set. The posterior distribution was also used to obtain a Bayes factor when compared to the null hypothesis that the categorical probabilities are uniformly distributed: $H_0 : p_1 = p_2 = \cdots = p_8$. These results were compared with a frequentist approach of hypothesis testing via a χ^2-test. Some deviations from uniformity were found with all three approaches, but ultimately the evidence was not strong enough to fully reject the null hypothesis.

Contents

1 Introduction

The MMORPG World of Warcraft has many forms of content, one of which includes "mythic+ dungeons". These dungeons feature a group of five players running a dungeon at an exponentially increasing difficulty for increased rewards. Mythic+ dungeons begin at a level of +2 and do not have a theoretical limit to how high they can scale, but will reach a practical ceiling due to the exponentially scaling nature. These dungeons also apply a timer to the run which is specific to the dungeon itself. Mythic+ dungeons also receive various "affixes" which rotate on a week-by-week basis. There are four tiers of affixes, and an affix from tier one is applied at +2, tier two is applied at +4, tier three is applied at +7, and tier four is applied at +10.

Every week, each player can receive a "keystone", or "key" for short. This keystone is selected from a pool of dungeons specific to the season that the game is currently in; the dungeon assigned to the key is the only dungeon it can be applied to. The level of the given key depends on the player's highest completed key within the season and their performance in the previous week. An example of a key can be seen in Figure 1

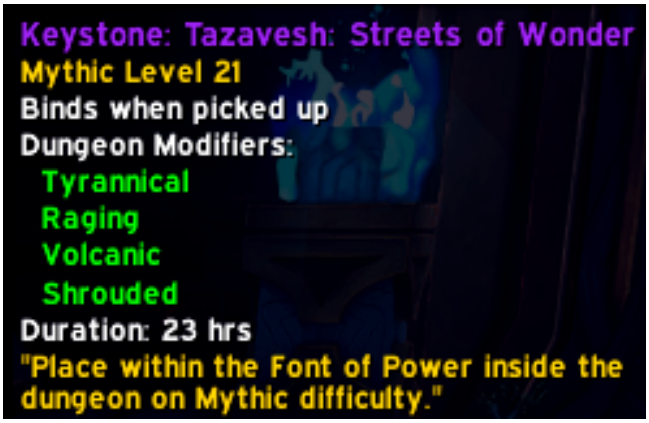

Figure 1: Example image of a keystone given to a player in game.

Keystones can be upgraded or downgraded based upon the group's performance in the dungeon. If the group succeeds in completing the dungeon within the allotted time, the keystone is upgraded by up to three levels, depending on the percentage of time remaining on the timer, and is assigned to a new dungeon from the other dungeons in the pool. If the group completes the dungeon after the timer has expired, the key is downgraded by one level and has a new dungeon assigned in the same way as a completion within time. If the group chooses to not complete the run, the keystone will be downgraded by one level but will remain at the same dungeon.

A natural thought that might arise when considering keystone distribution, is that the dungeon that a player is assigned should be uniformly chosen from the pool, i.e., with a dungeon pool of 8, each dungeon should have a 12.5% chance of being assigned for the first key and a roughly 14.3% chance of of being assigned for each key thereafter. But is this actually the case? After years of playing World of Warcraft, many players experience their keystone flipping back and forth between the same two dungeons, and some weeks it seems especially difficult to find someone who received a particular dungeon. While it could just

be a form of experience bias, there is also the possibility that keystones are not allocated equally. How, then, are dungeons actually distributed in World of Warcraft? Specifically, are the dungeons distributed uniformly?

The biggest reason one might be interested in this question is because of how this affects players in the game. Since many players are looking for a specific piece to upgrade their character with, or the player is looking to increase their overall dungeon score (which is a sum of scores for each dungeon individually), the relative frequencies at which keys appear could be important for proper time budgeting. If, for example, on the week where the "Fortified, Sanguine, Grievous" affix set appears there is an over abundance of GMBT keys but a shortage of STRT keys, but then on the "Fortified, Bursting, Storming" week the opposite case is true, then players who need GMBT and not STRT would look to play on the former week, rather than the latter if they cannot afford to spend time searching for a scarce key.

2 Data

Currently, World of Warcraft is in season 4 of its current expansion. In this season, there are eight dungeons within the dungeon pool: Tazavesh: Streets of Wonder (STRT), Tazavesh: So'leah's Gambit (GMBT), Operation: Mechagon - Junkyard (YARD), Operation: Mechagon - Workshop (WORK), Return to Karazhan: Upper (UPPR), Return to Karazhan: Lower (LOWR), Grimrail Depot (GD), and Iron Docks (ID). The affixes that were present during data collection include fortified, tyrannical, bolstering, bursting, inspiring, raging, sanguine, spiteful, explosive, grievous, necrotic, quaking, storming, and volcanic.

To determine how the dungeons are distributed, data was collected from various players over the course of six weeks. The first key each player receives within a week is assumed to be uniformly distributed across all eight dungeons in the pool. However, subsequent keys have a dependence on what the previous key was, e.g., a STRT key will not turn into a STRT key if completed, and hence has a pool of the eight dungeons minus whichever dungeon the key was. To simplify the model, and by extension computations, only the first key a player received each week was collected. When collected, the dungeon location, the keystone level, and the affixes for the week were recorded. Since the affixes rotate weekly, and do not vary key by key, all weekly affixes were included on all key levels, even on keys too low to receive a given affix. Similarly, the tier four affix was ignored as it rotates on a seasonal basis and is hence present on all keys of the appropriate level. As a note, players were able to submit more than one keystone per week provided the keystone was from a separate character, as characters within the game are, for all intents and purposes, different players, and different characters are identical to different players within the scope of this topic.

The collected data can be seen in Figures (2)-(21), broken down by all data, affix sets, individual affixes, and key levels. Some figures were ignored, e.g., single affix distributions for all except the fortified affix and the tyrannical affix, as the information was redundant with other figures due to the data collection only occurring over a six week period rather than a whole cycle of twelve weeks. The twelve week cycle would allow affixes to be repeated in other combinations would hence be more relevant. Similarly, only key levels where at least 16 (twice the dungeon pool) observations were recorded are included.

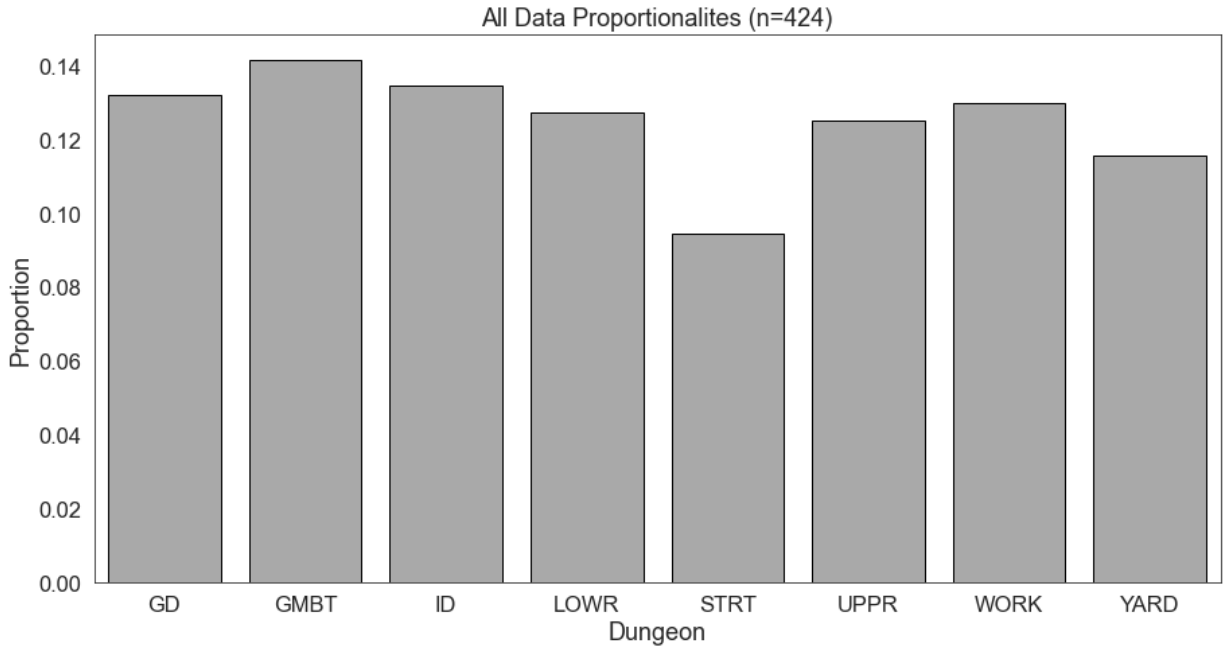

Figure 2: Bar plot of the proportions of occurrence for each dungeon in the collected data.

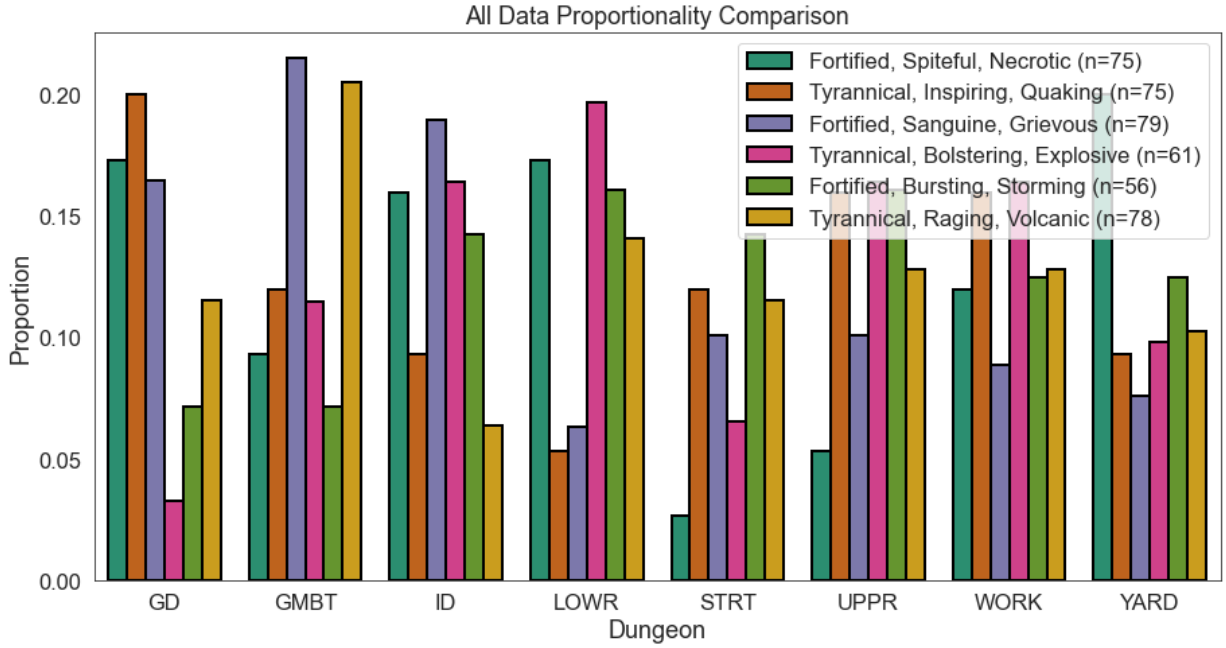

Figure 3: Side-by-side comparison of the distributions for each dungeon per week in the collected data.

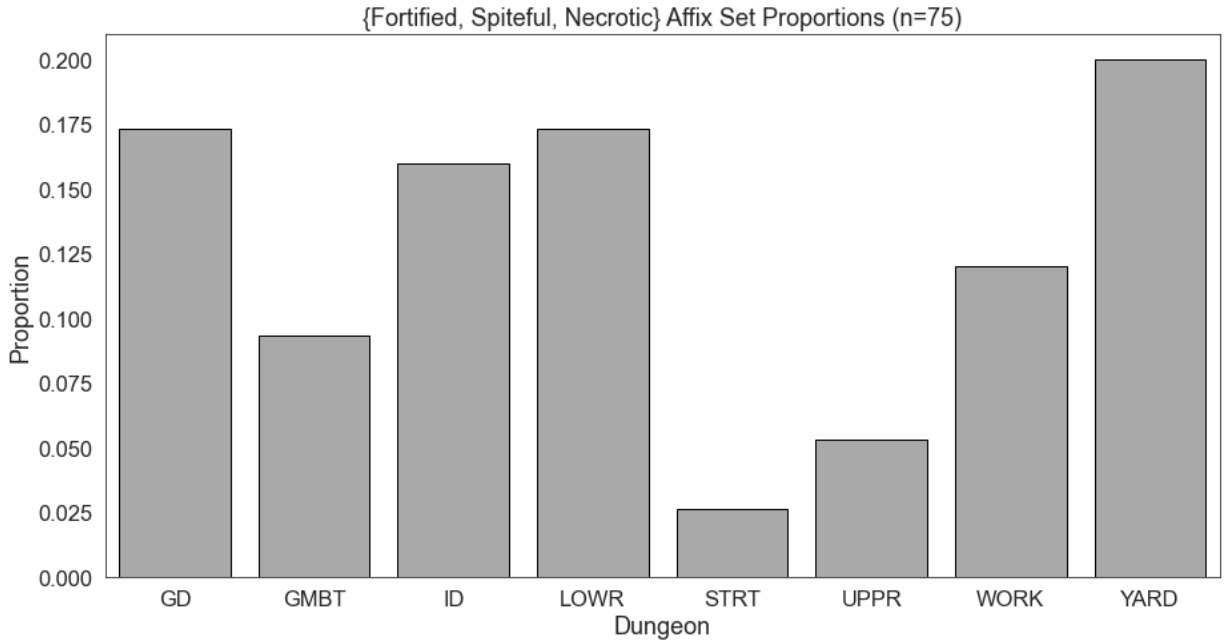

Figure 4: Bar plot of the proportions of occurrence for dungeons with the fortified, spiteful, and necrotic affix set in the collected data.

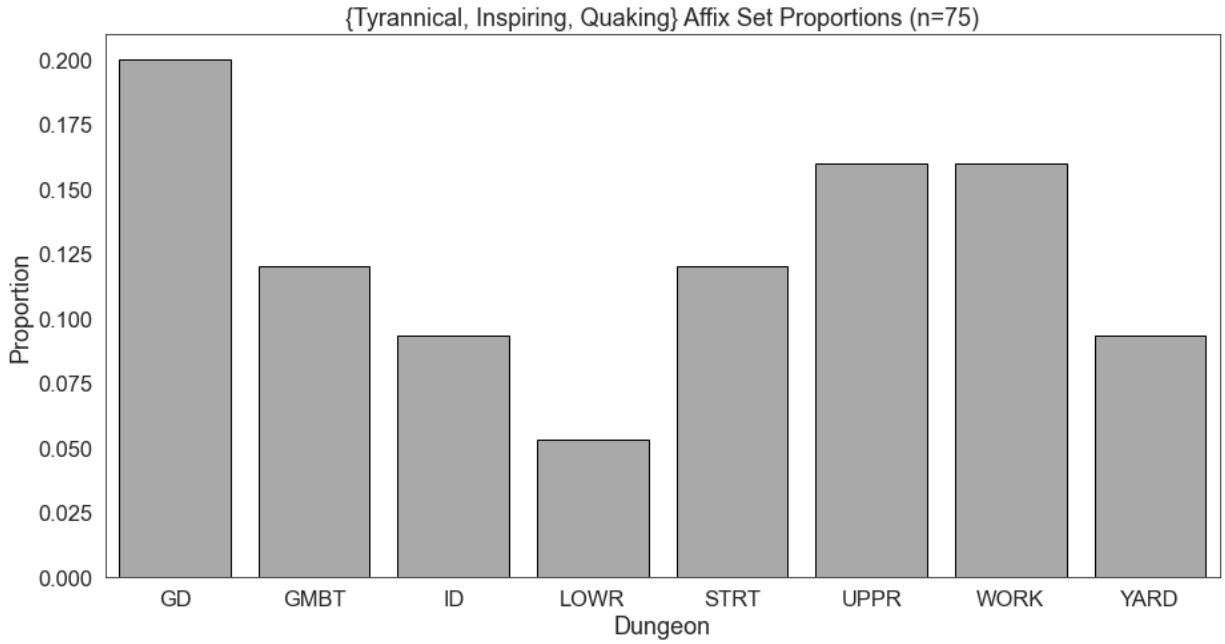

Figure 5: Bar plot of the proportions of occurrence for dungeons with the tyrannical, inspiring, and quaking affix set in the collected data.

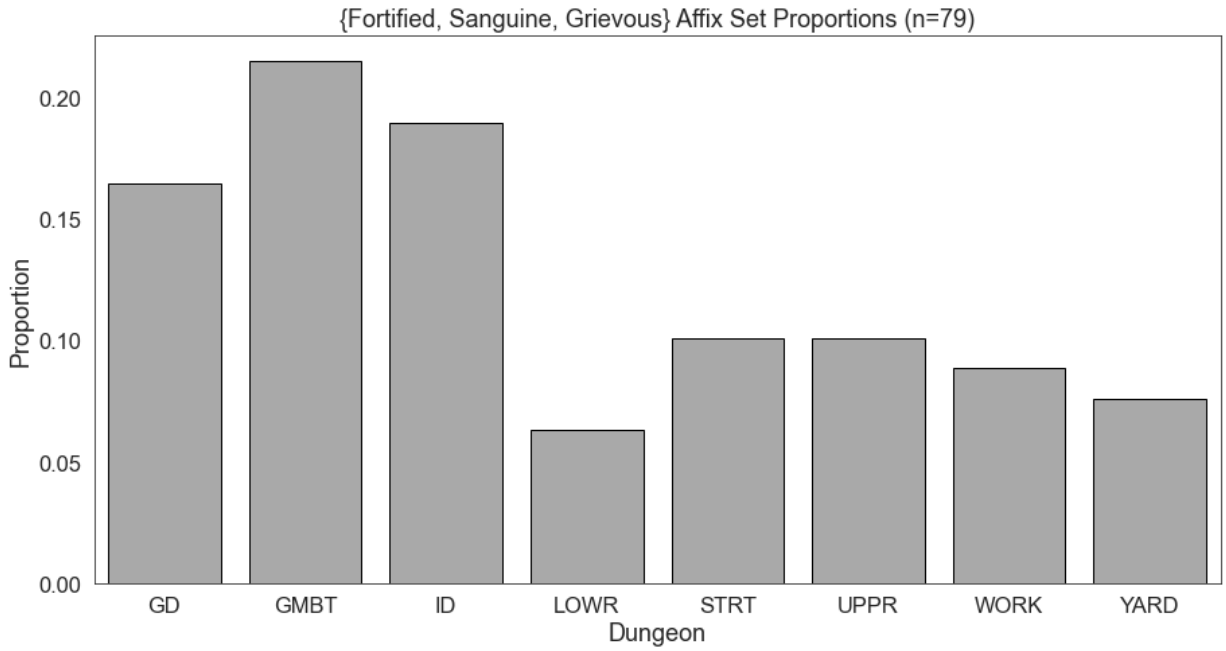

Figure 6: Bar plot of the proportions of occurrence for dungeons with the fortified, sanguine, and grievous affix set in the collected data.

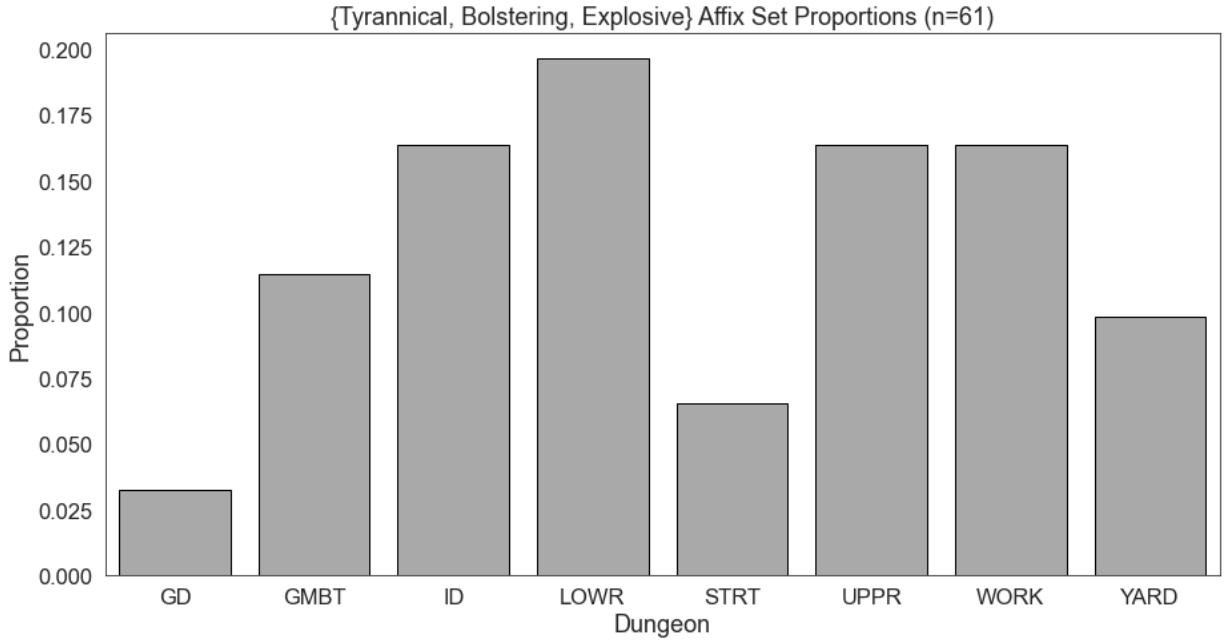

Figure 7: Bar plot of the proportions of occurrence for dungeons with the tyrannical, bolstering, and explosive affix set in the collected data.

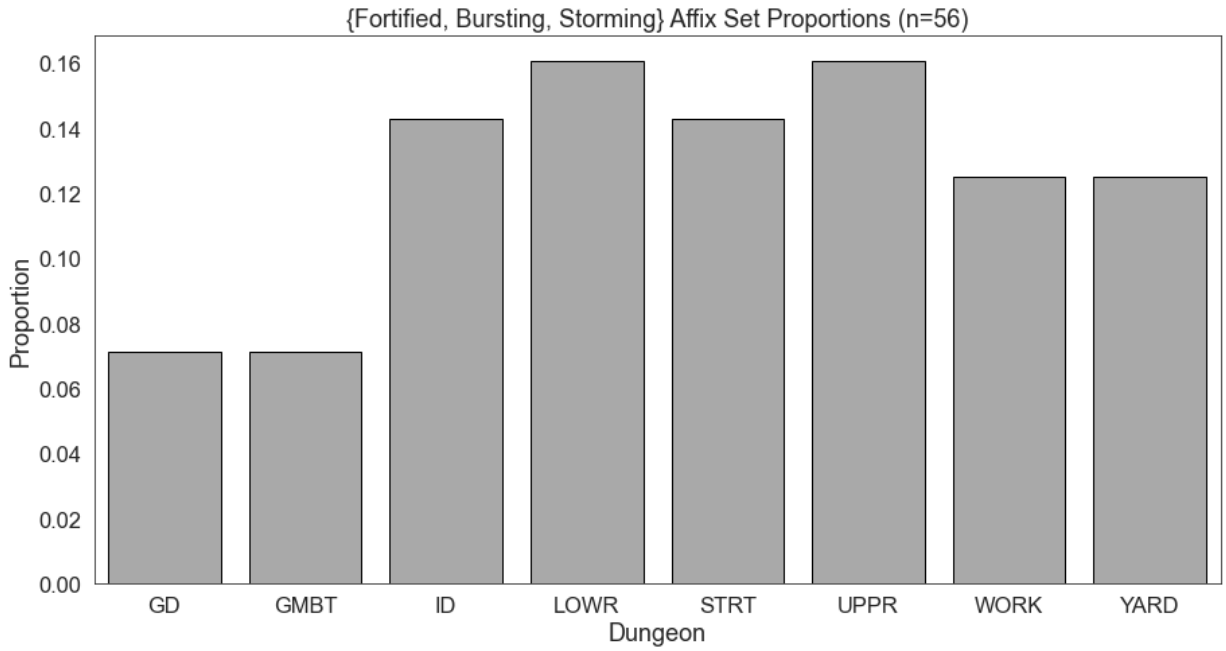

Figure 8: Bar plot of the proportions of occurrence for dungeons with the fortified, bursting, and storming affix set in the collected data.

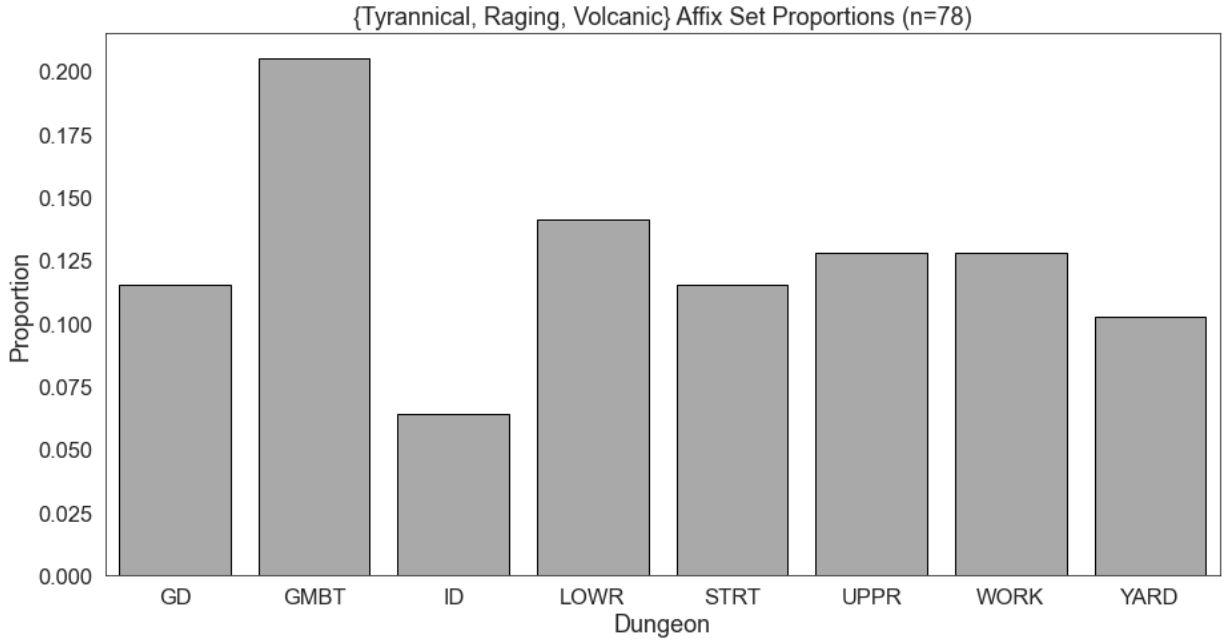

Figure 9: Bar plot of the proportions of occurrence for dungeons with the tyrannical, raging, and volcanic affix set in the collected data.

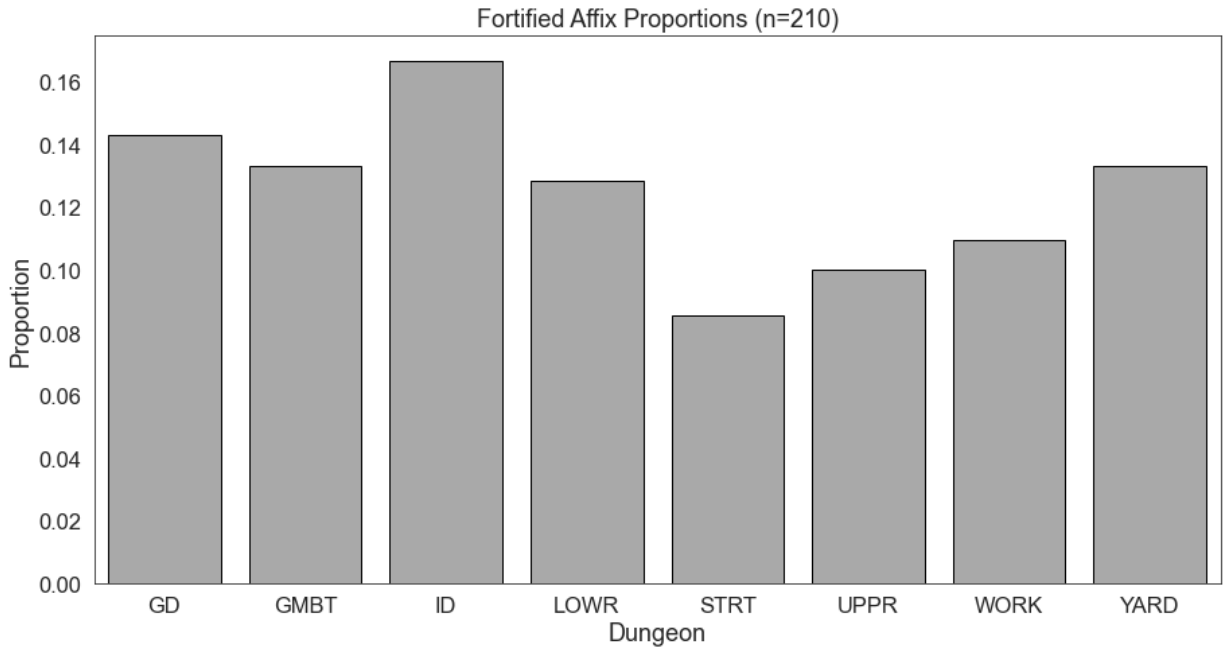

Figure 10: Bar plot of the proportions of occurrence for dungeons with the fortified affix in the collected data.

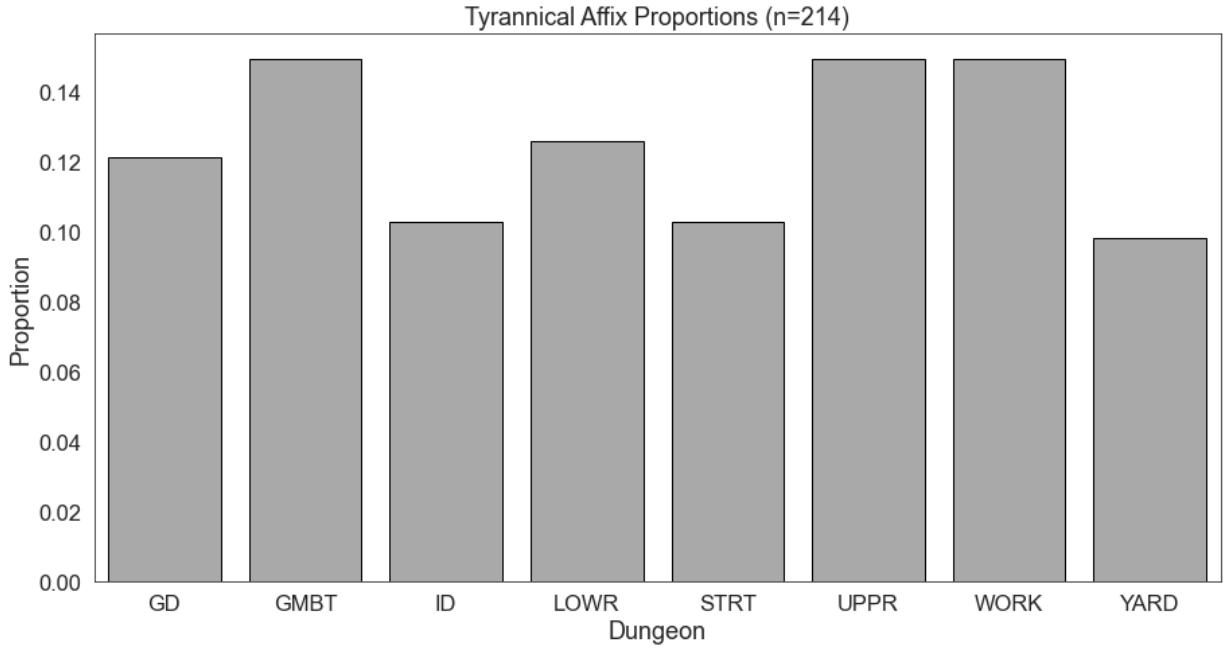

Figure 11: Bar plot of the proportions of occurrence for dungeons with the tyrannical affix in the collected data.

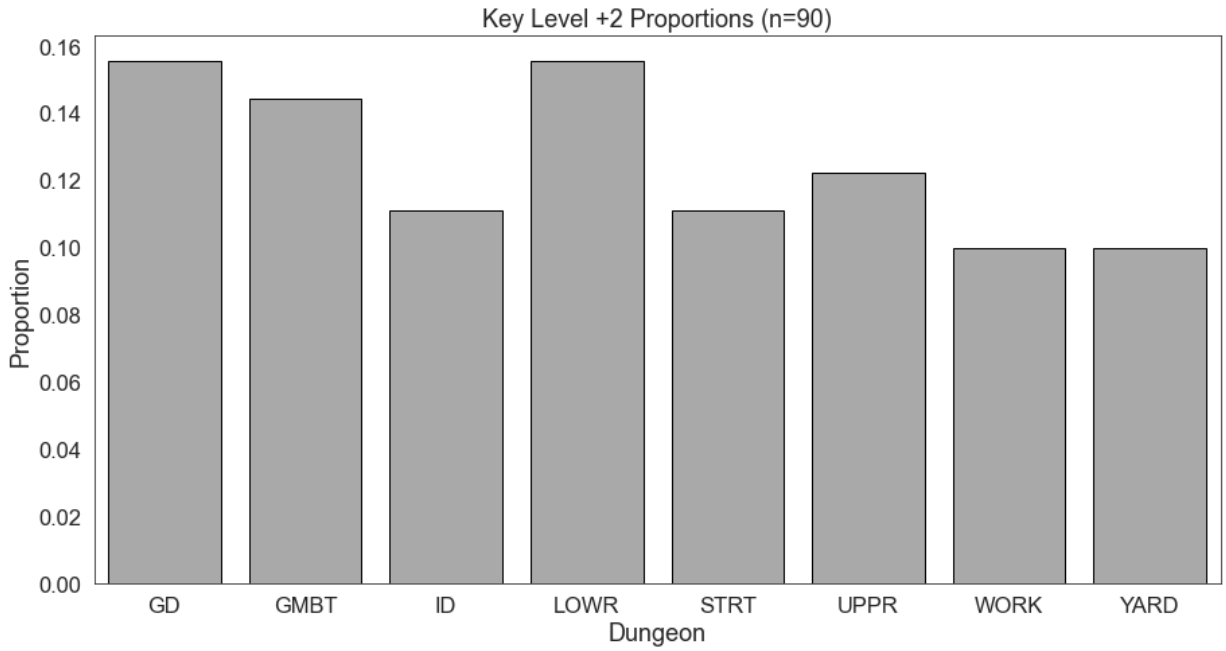

Figure 12: Bar plot of the proportions of occurrence for dungeons at the +2 level in the collected data.

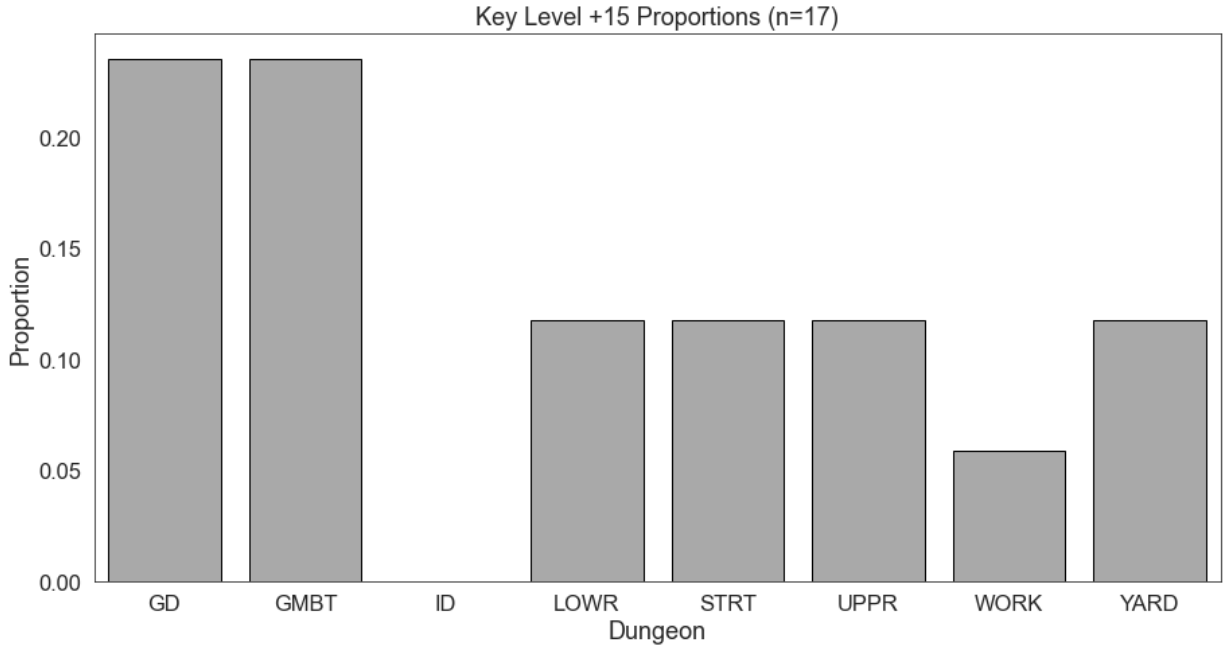

Figure 13: Bar plot of the proportions of occurrence for dungeons at the +15 level in the collected data.

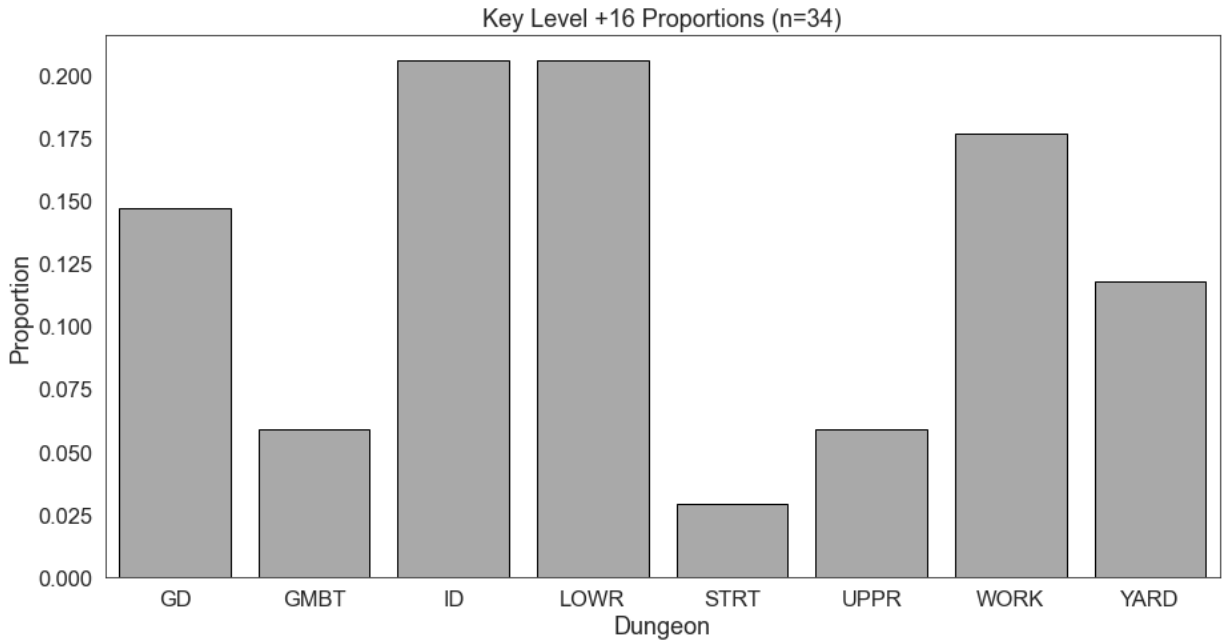

Figure 14: Bar plot of the proportions of occurrence for dungeons at the +16 level in the collected data.

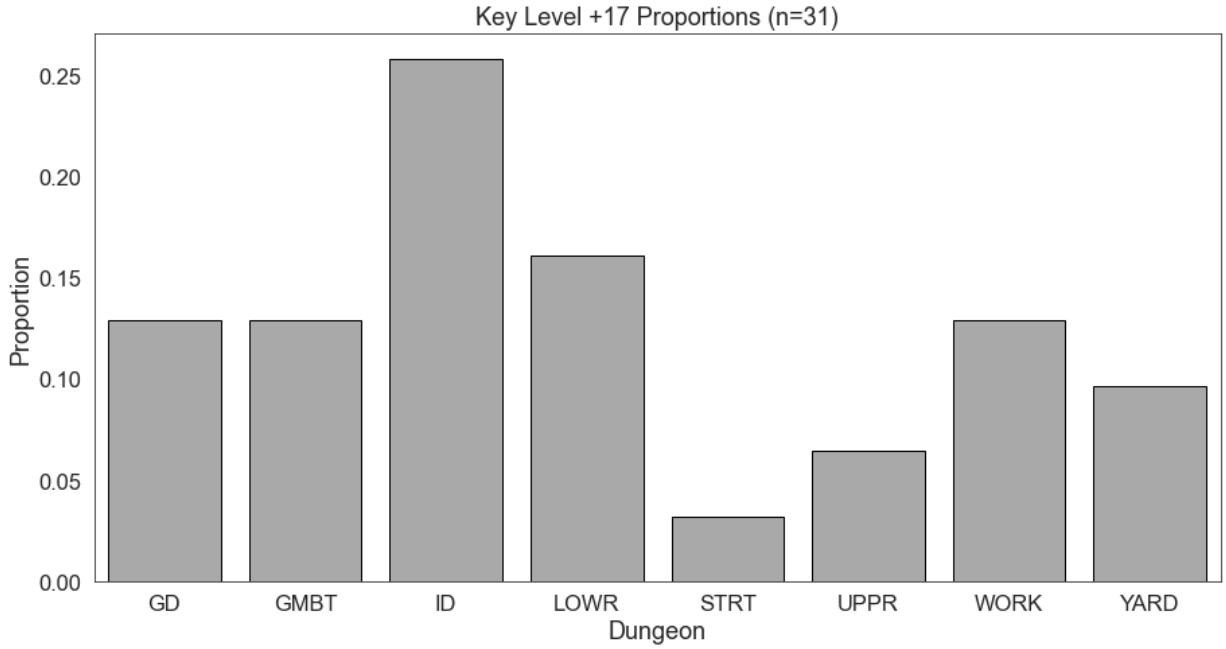

Figure 15: Bar plot of the proportions of occurrence for dungeons at the +17 level in the collected data.

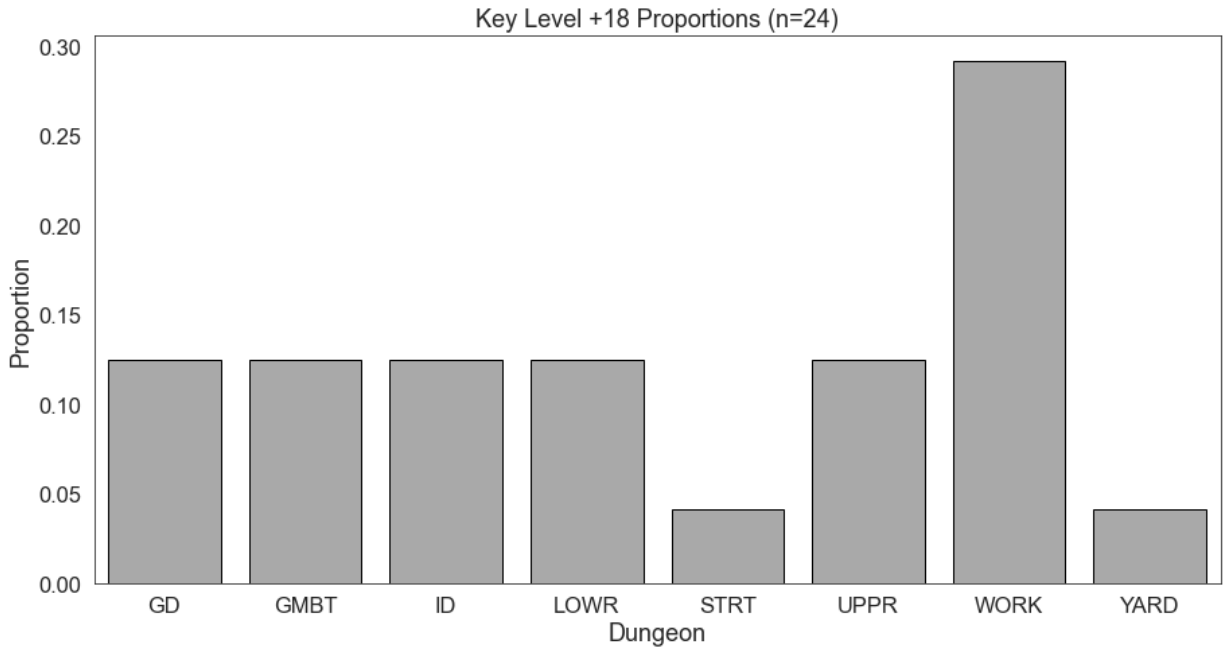

Figure 16: Bar plot of the proportions of occurrence for dungeons at the +18 level in the collected data.

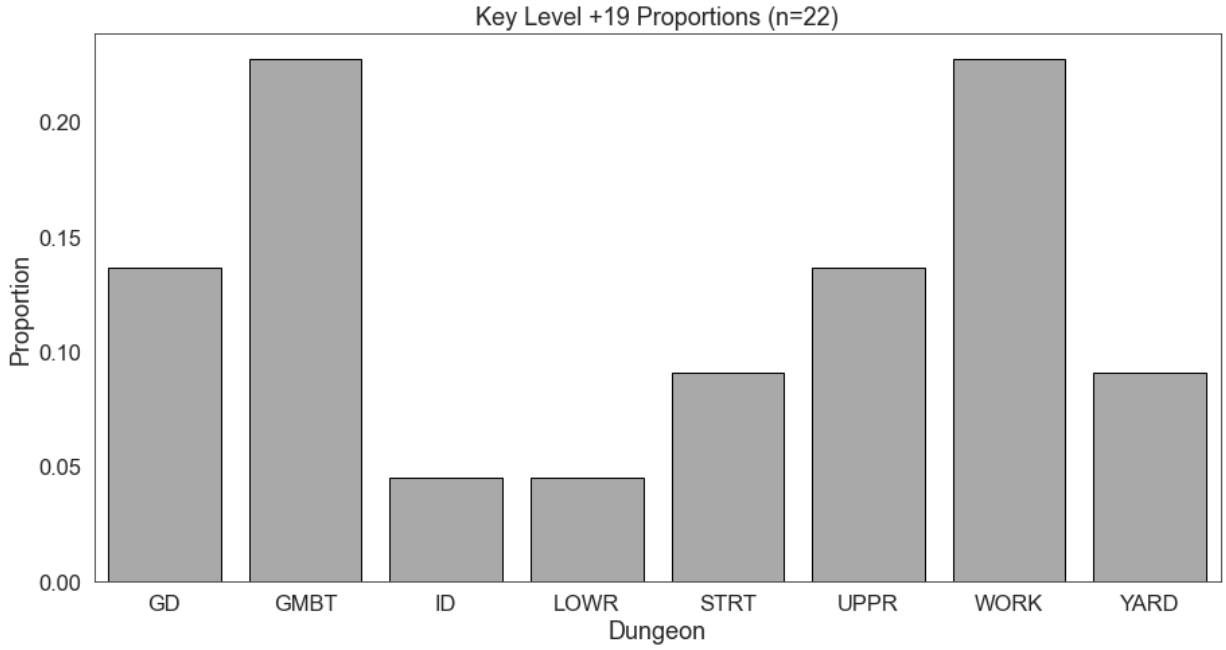

Figure 17: Bar plot of the proportions of occurrence for dungeons at the +19 level in the collected data.

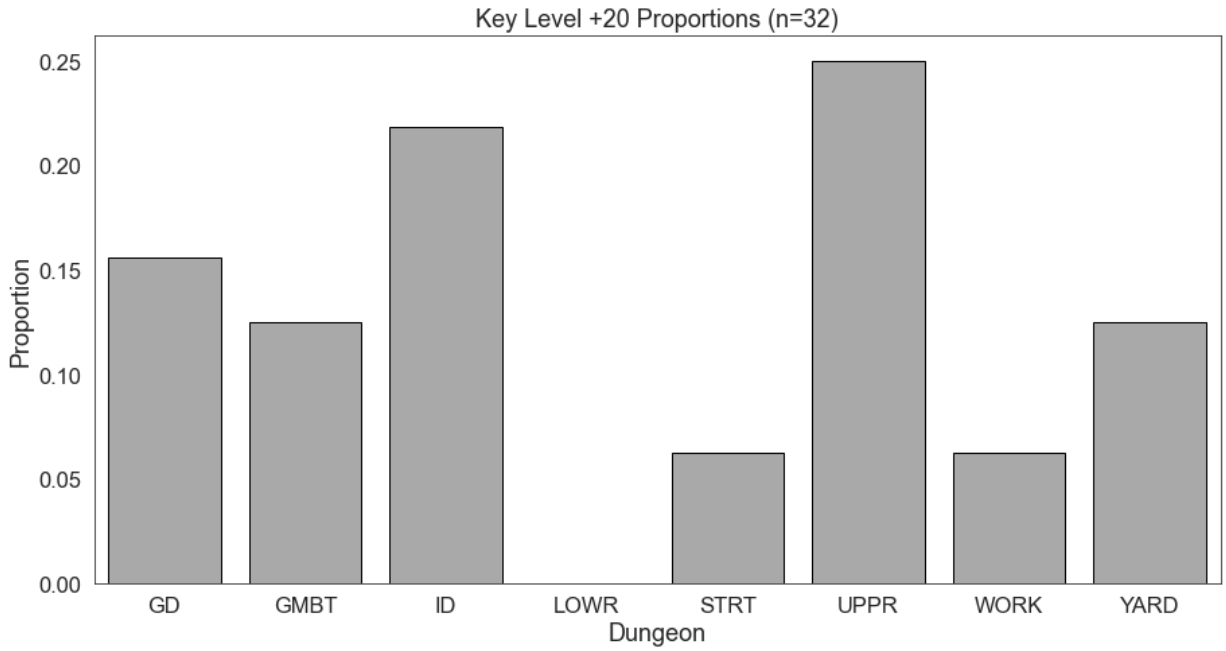

Figure 18: Bar plot of the proportions of occurrence for dungeons at the +20 level in the collected data.

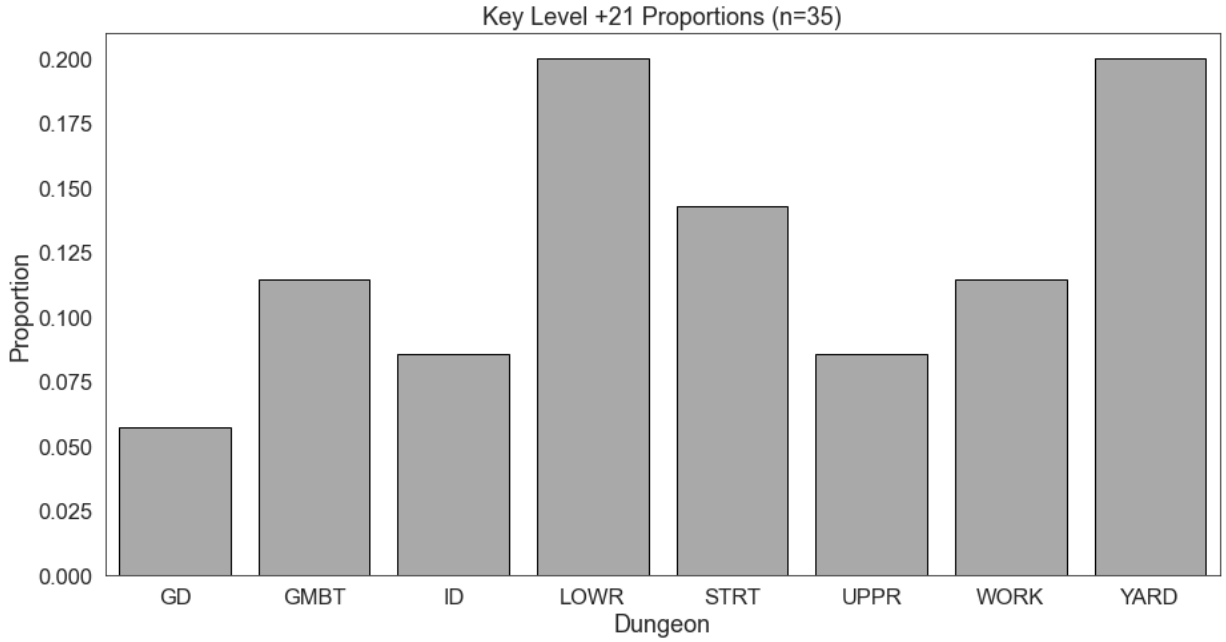

Figure 19: Bar plot of the proportions of occurrence for dungeons at the +21 level in the collected data.

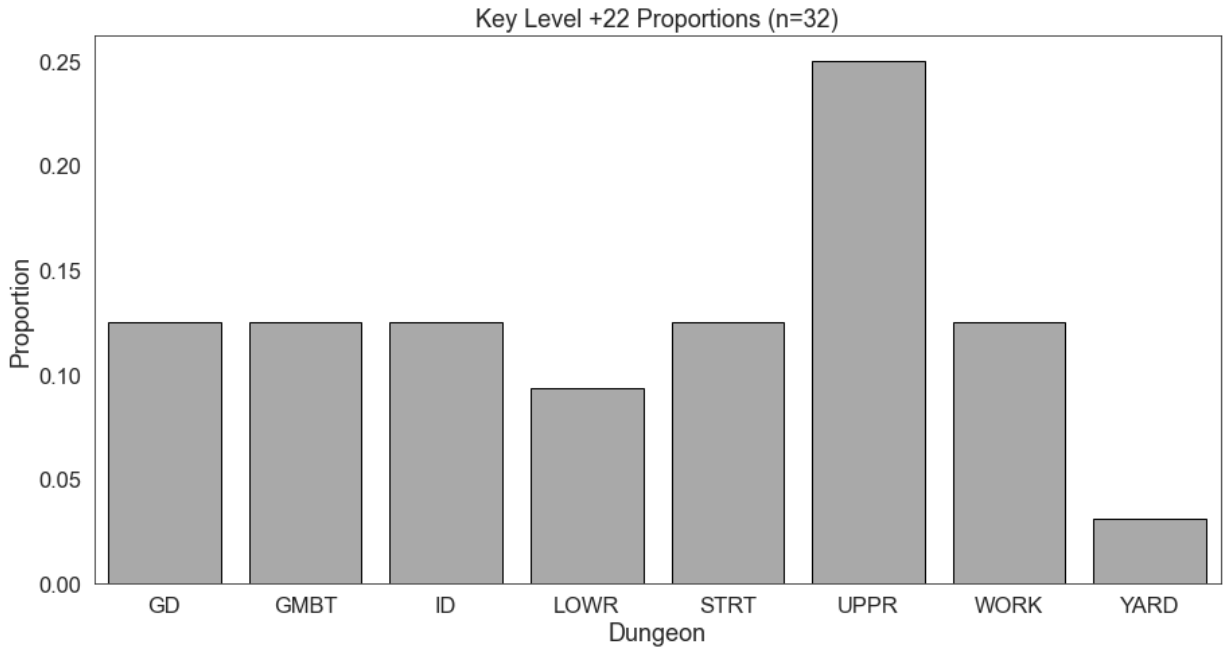

Figure 20: Bar plot of the proportions of occurrence for dungeons at the +22 level in the collected data.

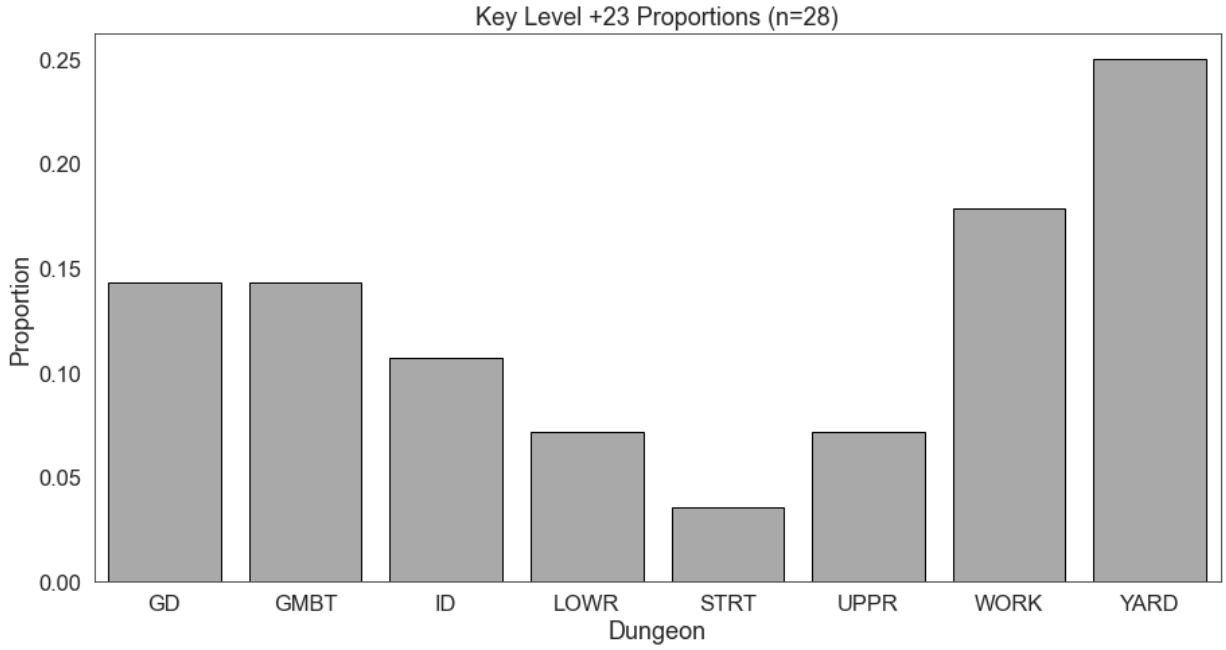

Figure 21: Bar plot of the proportions of occurrence for dungeons at the +23 level in the collected data.

3 Model

The collection of all keystones $\mathbf{Y}$, meeting a set of conditions can be recognized as coming from a multinomial distribution with a number of categories, k equal to the size of the dungeon pool ($k = 8$ at the time of this analysis) and a number of trials, n, equal to the number of keystones under said conditions where each Y_i is the number of trials in category $i \in \{1, \ldots, k\}$. To simplify computations, the prior distribution used will be a Dirichlet distribution, as the Dirichlet distribution is a conjugate prior of the multinomial distribution. Now it is fair to assume that dungeons would be uniformly distributed under any circumstance, and as such the Dirichlet prior will have its sole parameter be a k-dimensional vector where $\alpha_i = z \quad \forall i \in \{1, ..., k\}$ such that $z \in \mathbb{R} > 0$. Larger values of z indicate a stronger assertion that the prior assumption that the probabilities of each dungeon occurring are equal. In this case, the value of $z = 1$ is taken in order to use a weakly-informative uniform prior.

Then,

$$Y \sim \text{Multinomial}(Y|n, \mathbf{p})$$

$$\mathbf{p} \sim \text{Dirichlet}(\mathbf{p}|\boldsymbol{\alpha})$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1, & \ldots, & Y_k \end{pmatrix}$$

$$\mathbf{p} = \begin{pmatrix} p_1, & \ldots, & p_k \end{pmatrix}$$

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1, & \ldots, & \alpha_k \end{pmatrix}.$$

Hence,

$$\Pr(\mathbf{p}|\mathbf{Y}) \propto \text{Dirichlet}(\mathbf{p}|\boldsymbol{\alpha}) \prod_{i=1}^{N} \text{Multinomial}(\mathbf{Y}|\mathbf{p})$$

$$\Pr(\mathbf{p}|\mathbf{Y}) \propto \left(\prod_{j=1}^{k} p_j^{\alpha_j - 1} \right) \left(\prod_{j=1}^{k} p_j^{Y_j} \right)$$

$$\Pr(\mathbf{p}|\mathbf{Y}) \propto \prod_{j=1}^{k} p_j^{\alpha_j + Y_j - 1}$$

Therefore,

$$\mathbf{p}|\mathbf{Y} \sim \text{Dirichlet}(\mathbf{p}|\boldsymbol{\alpha'})$$

where

$$\boldsymbol{\alpha'} = \begin{pmatrix} \alpha_1 + Y_1, & \ldots, & \alpha_k + Y_k \end{pmatrix} = \begin{pmatrix} 1 + Y_1, & \ldots, & 1 + Y_k \end{pmatrix}.$$

4 Results

Initially, all data was analyzed together in one lump data set without conditioning on affixes or key level. Then the data was broken down by getting conditioned on weekly affix set, then by individual affixes, and then again by key level. Some combinations were left out in accordance with the reasoning mentioned in Section (2). The posterior distributions shown in Figures (22)-(40) were created using 10,000 samples of the probability vector, $\mathbf{p}$ from the posterior distributions, and then each p_i was plotted independently due to dimensional constraints.

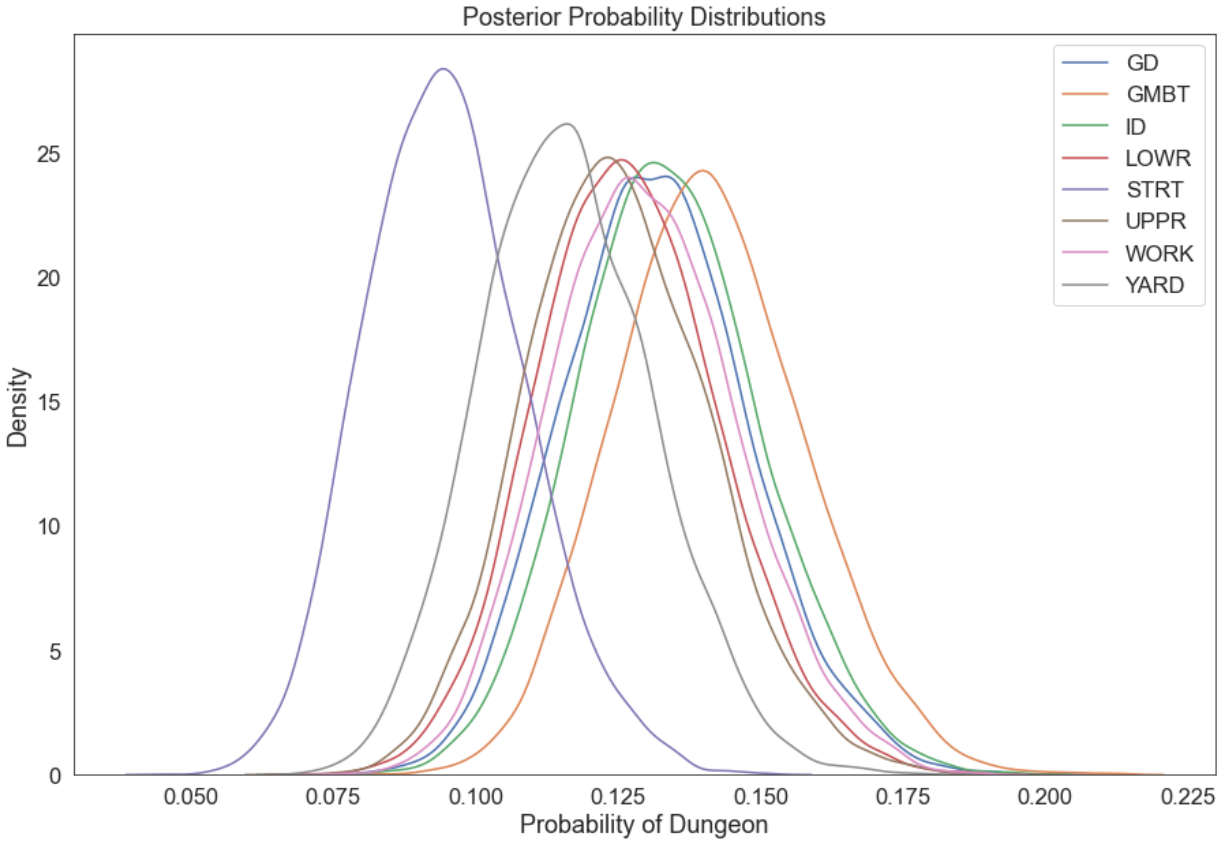

Figure 22: Posterior distributions for the probabilities of each dungeon for all data.

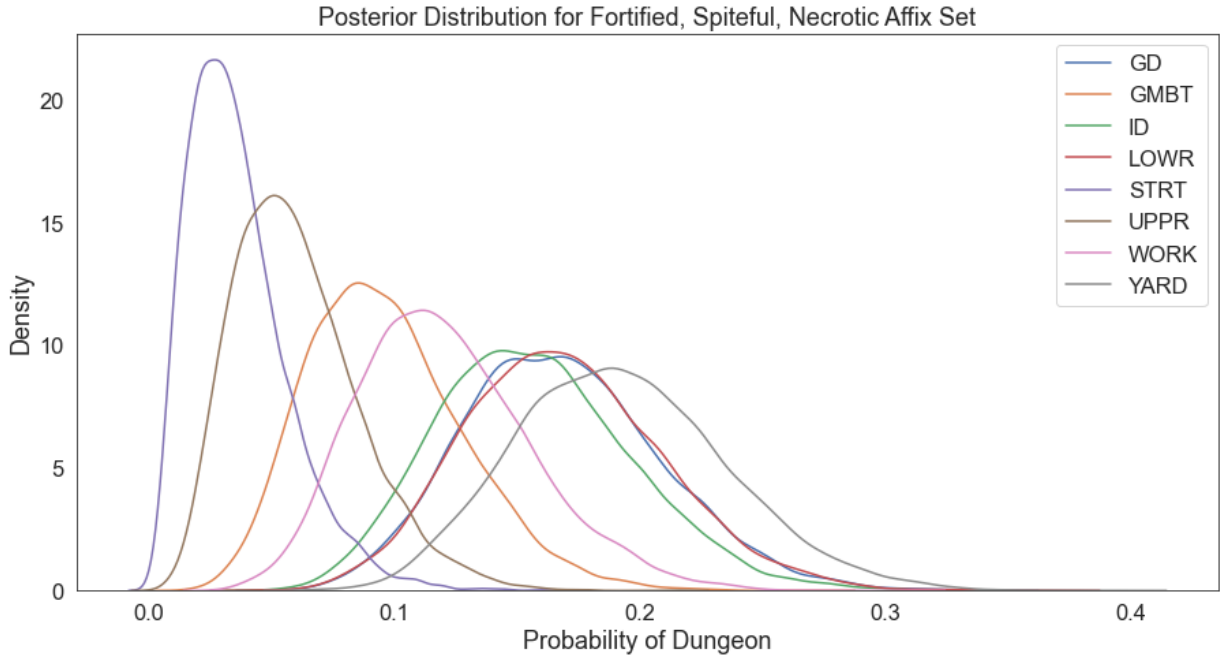

Figure 23: Posterior distributions for the probabilities of each dungeon with the fortified, spiteful, and necrotic affix set.

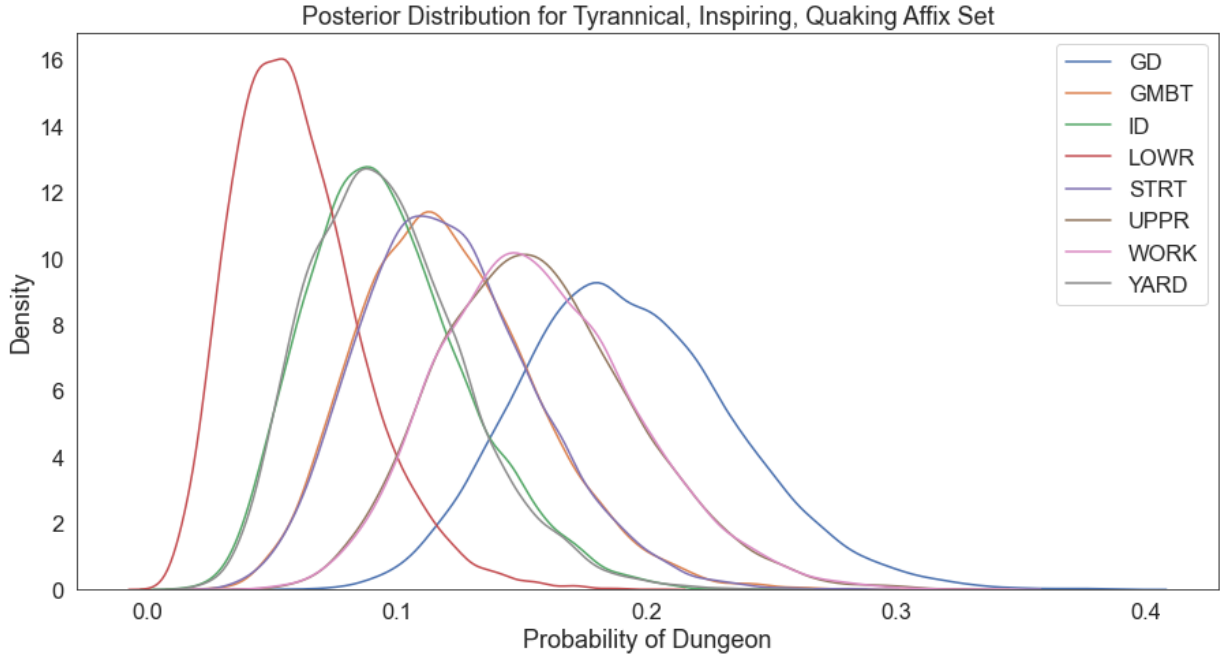

Figure 24: Posterior distributions for the probabilities of each dungeon with the tyrannical, inspiring, and quaking affix set.

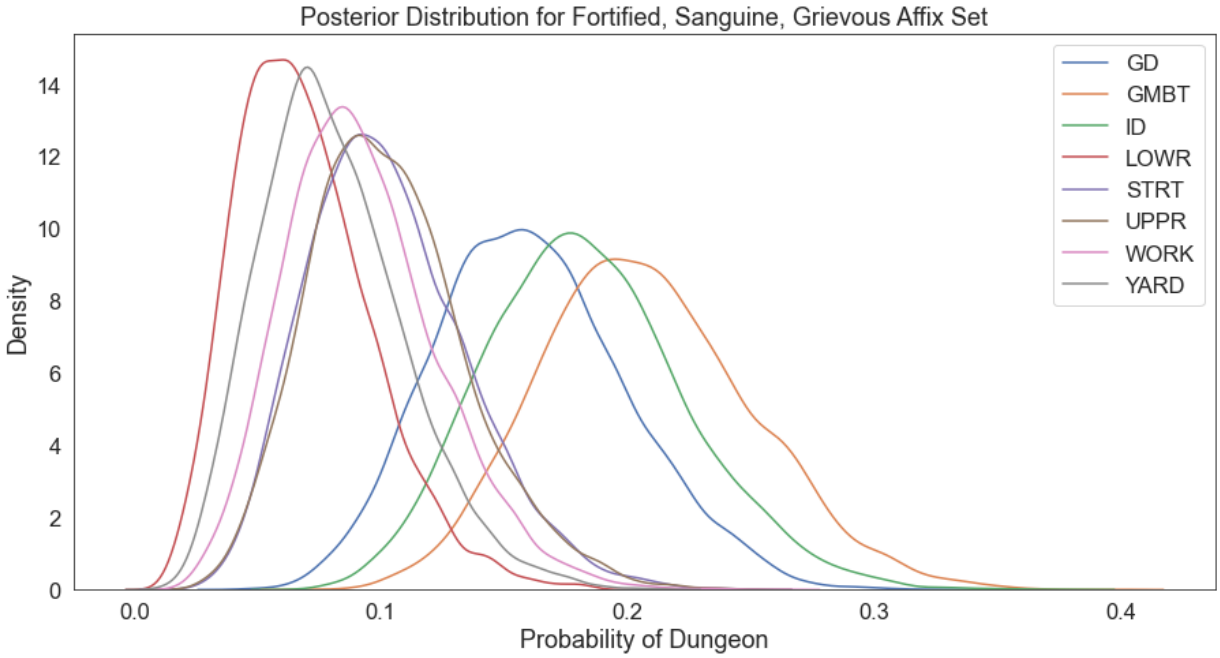

Figure 25: Posterior distributions for the probabilities of each dungeon with the fortified, sanguine, and grievous affix set.

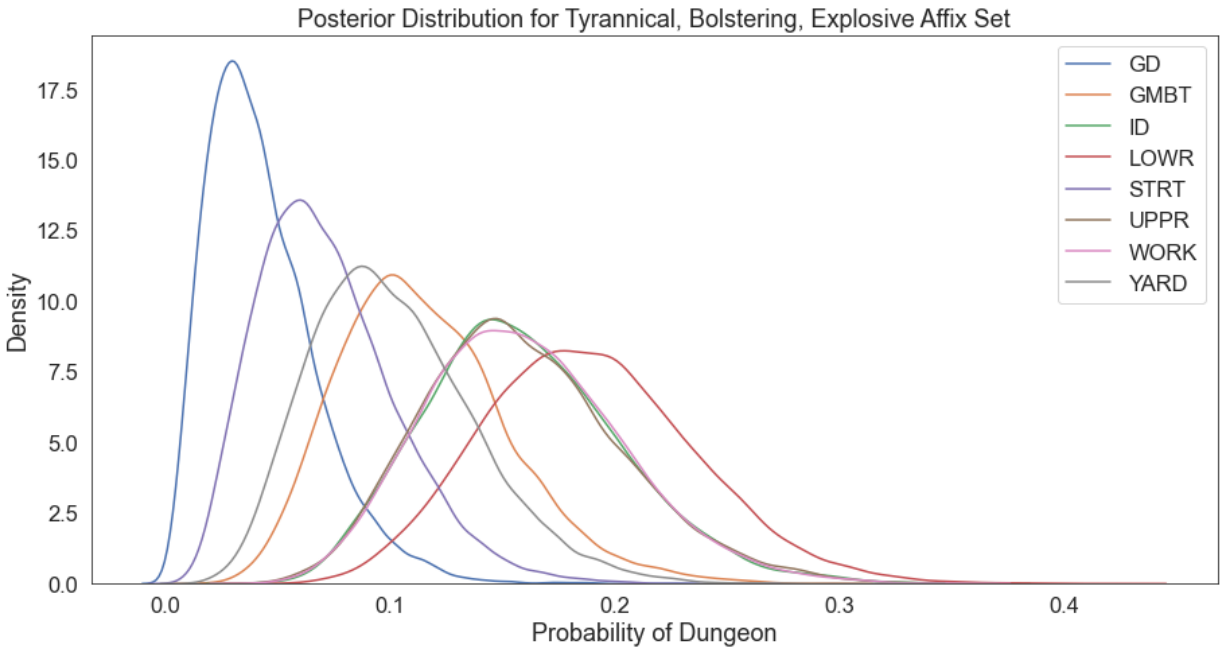

Figure 26: Posterior distributions for the probabilities of each dungeon with the tyrannical, bolstering, and explosive affix set.

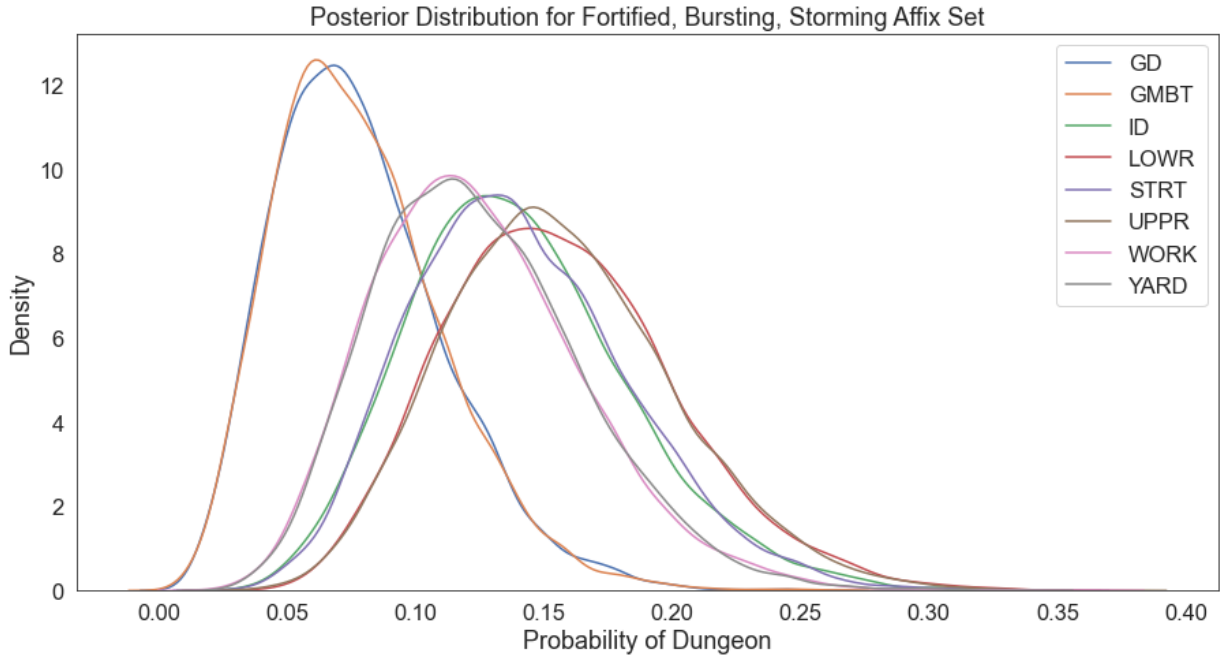

Figure 27: Posterior distributions for the probabilities of each dungeon with the fortified, bursting, and storming affix set.

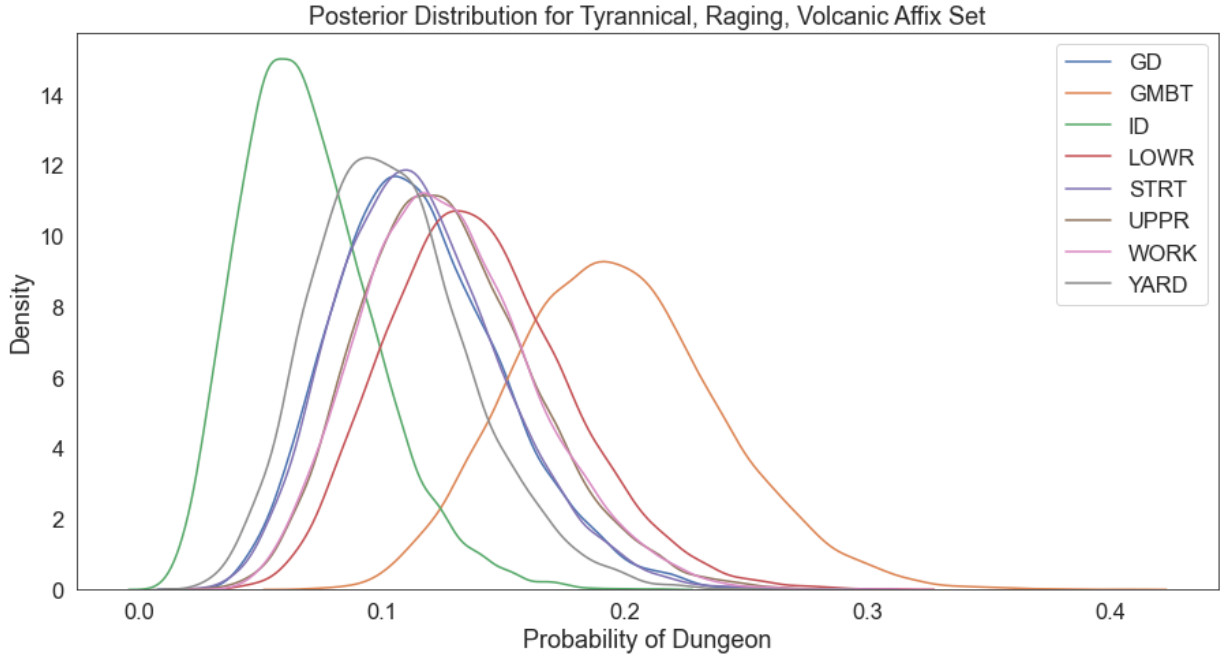

Figure 28: Posterior distributions for the probabilities of each dungeon with the tyrannical, raging, and volcanic affix set.

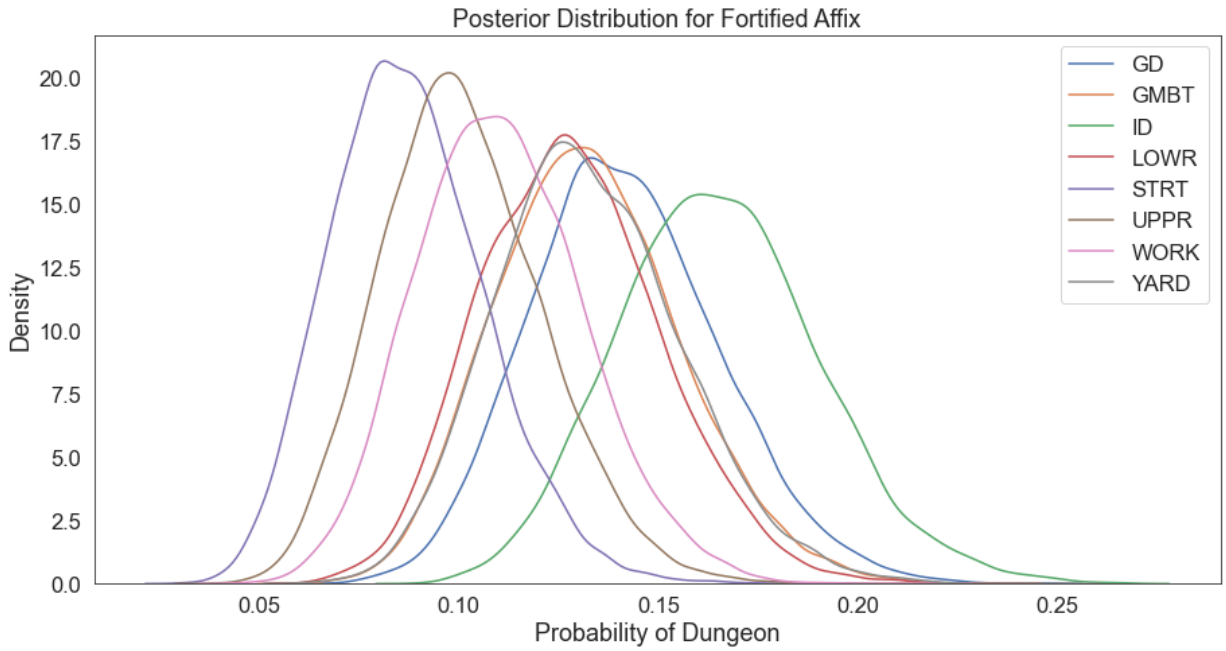

Figure 29: Posterior distributions for the probabilities of each dungeon with the fortified affix.

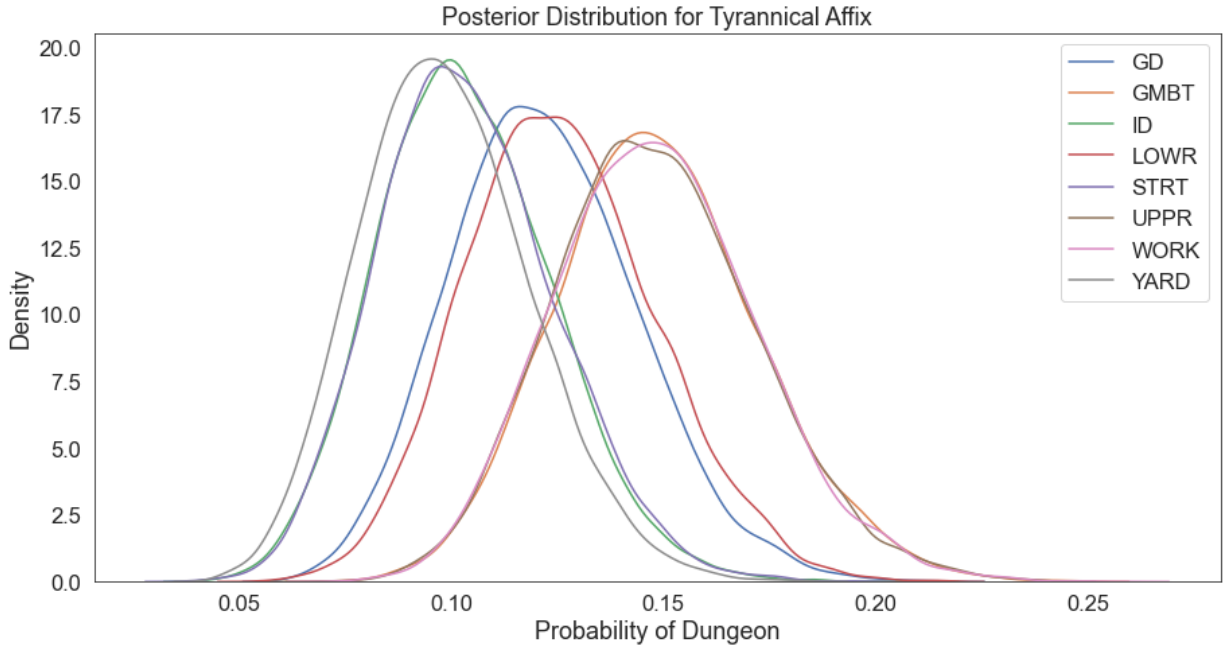

Figure 30: Posterior distributions for the probabilities of each dungeon with the tyrannical affix.

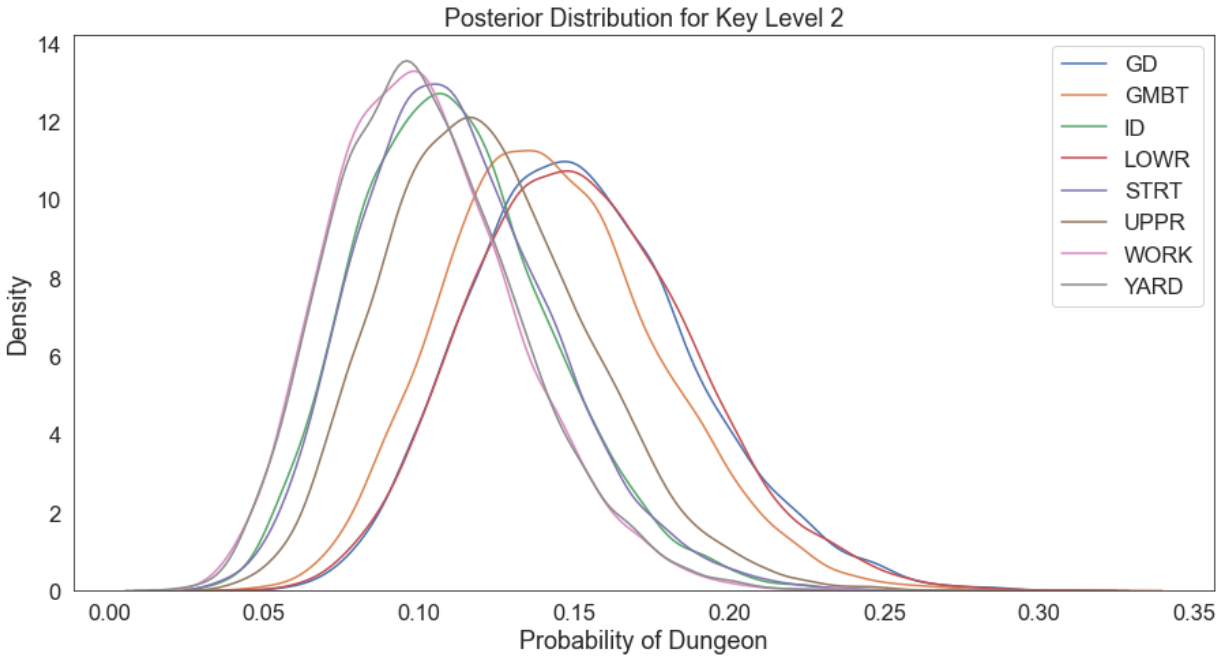

Figure 31: Posterior distributions for the probabilities of each dungeon at the +2 level.

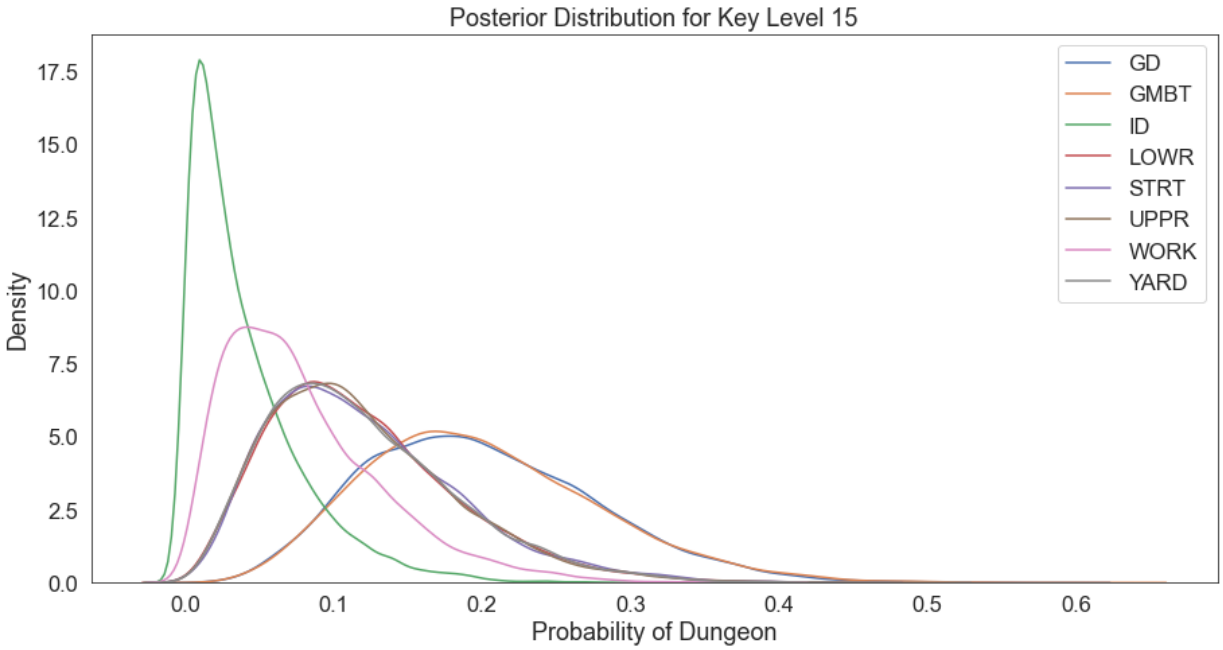

Figure 32: Posterior distributions for the probabilities of each dungeon at the +15 level.

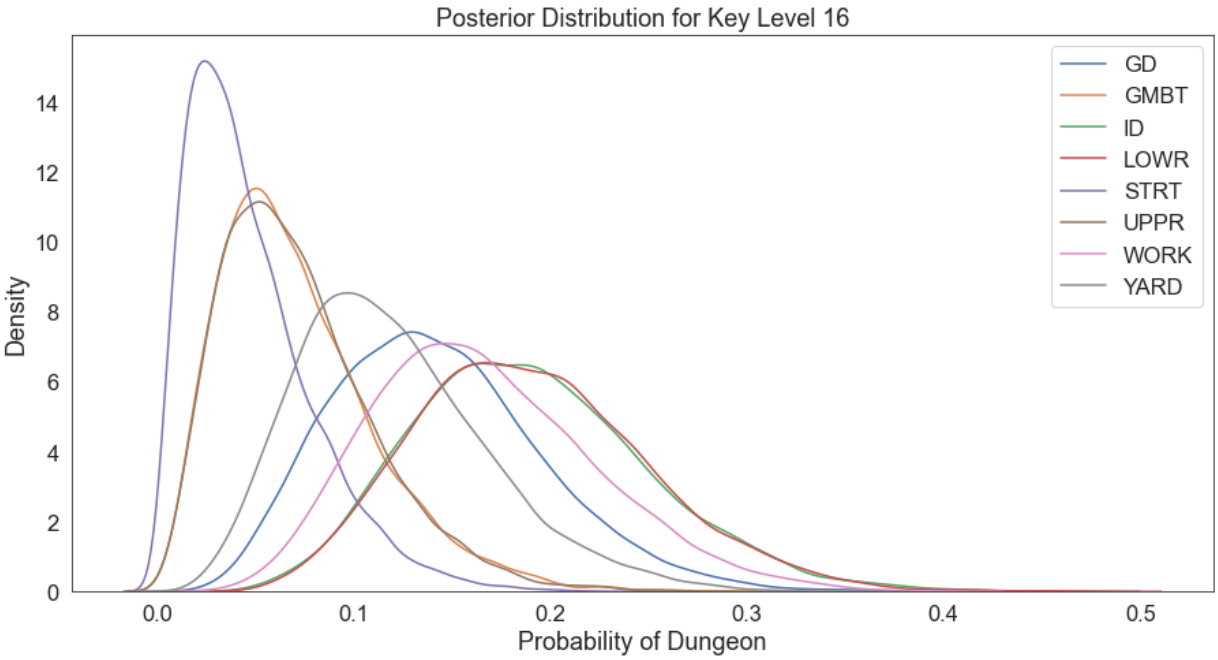

Figure 33: Posterior distributions for the probabilities of each dungeon at the +16 level.

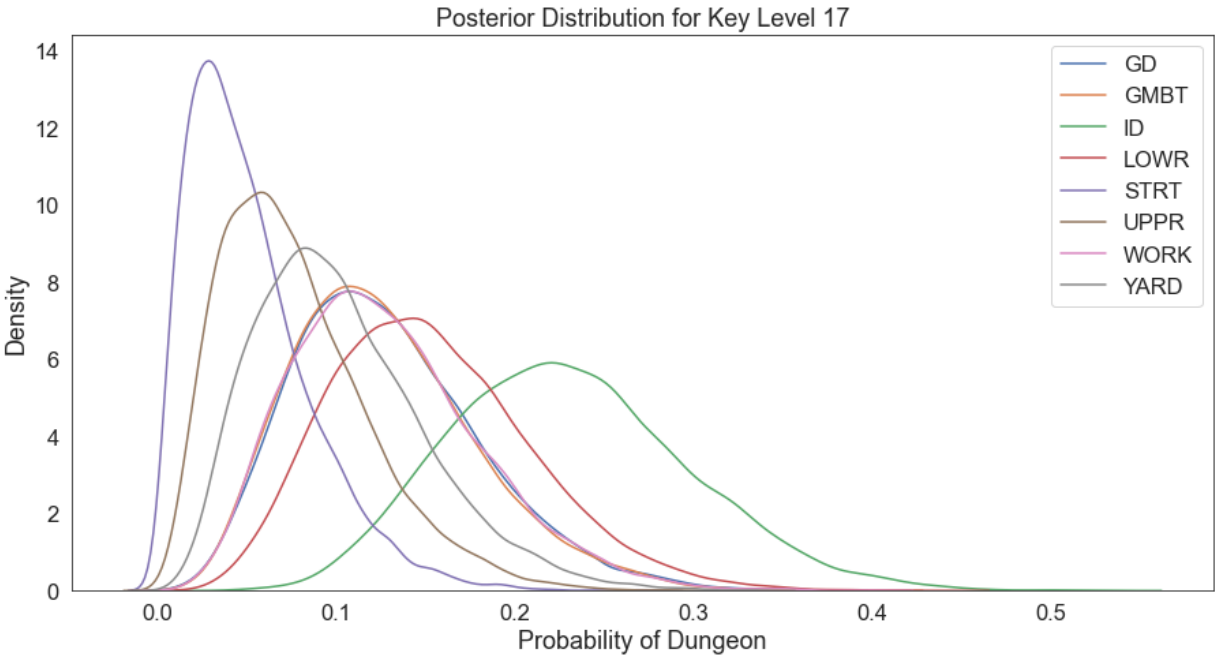

Figure 34: Posterior distributions for the probabilities of each dungeon at the +17 level.

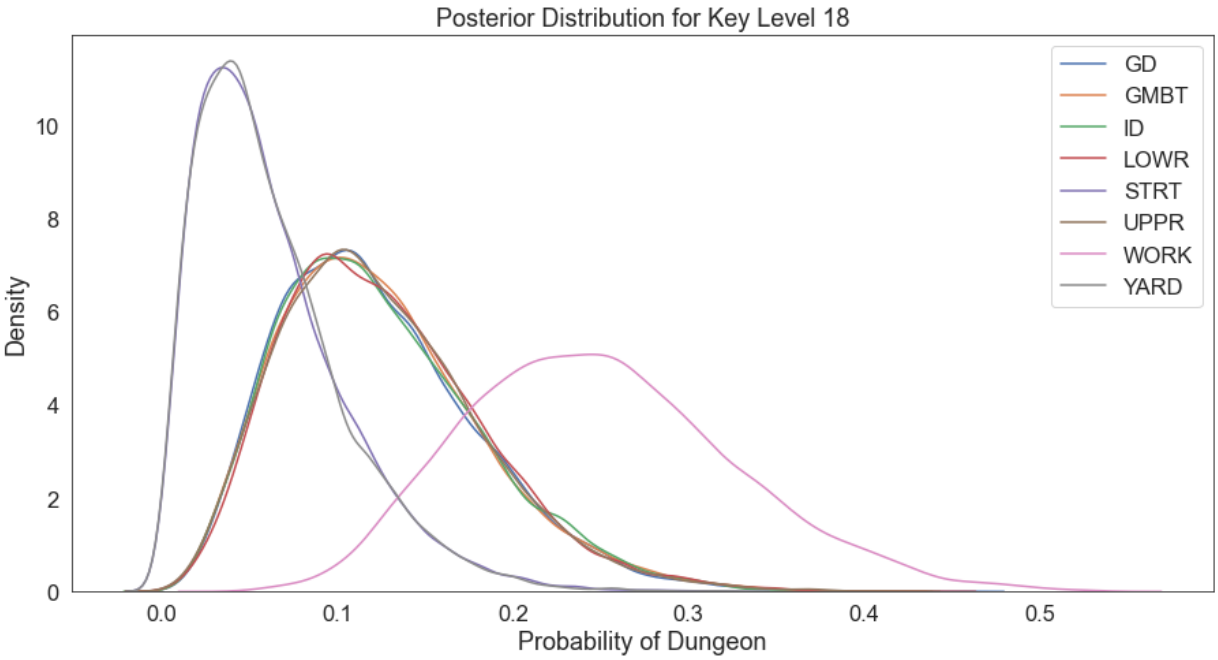

Figure 35: Posterior distributions for the probabilities of each dungeon at the +18 level.

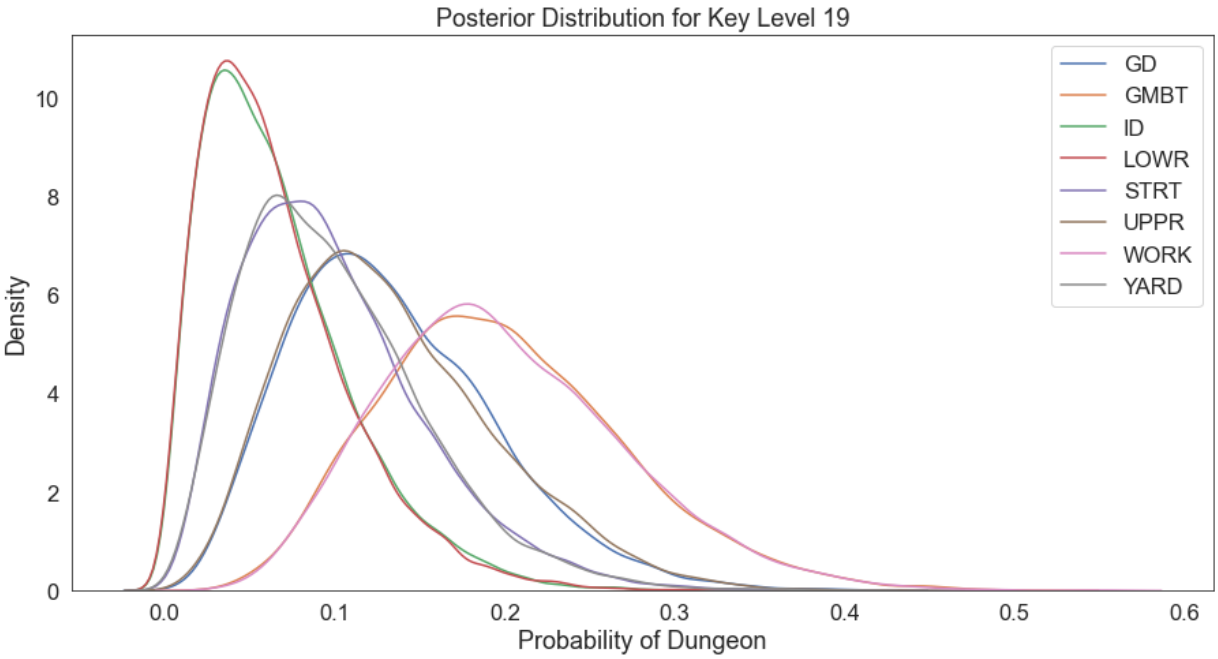

Figure 36: Posterior distributions for the probabilities of each dungeon at the +19 level.

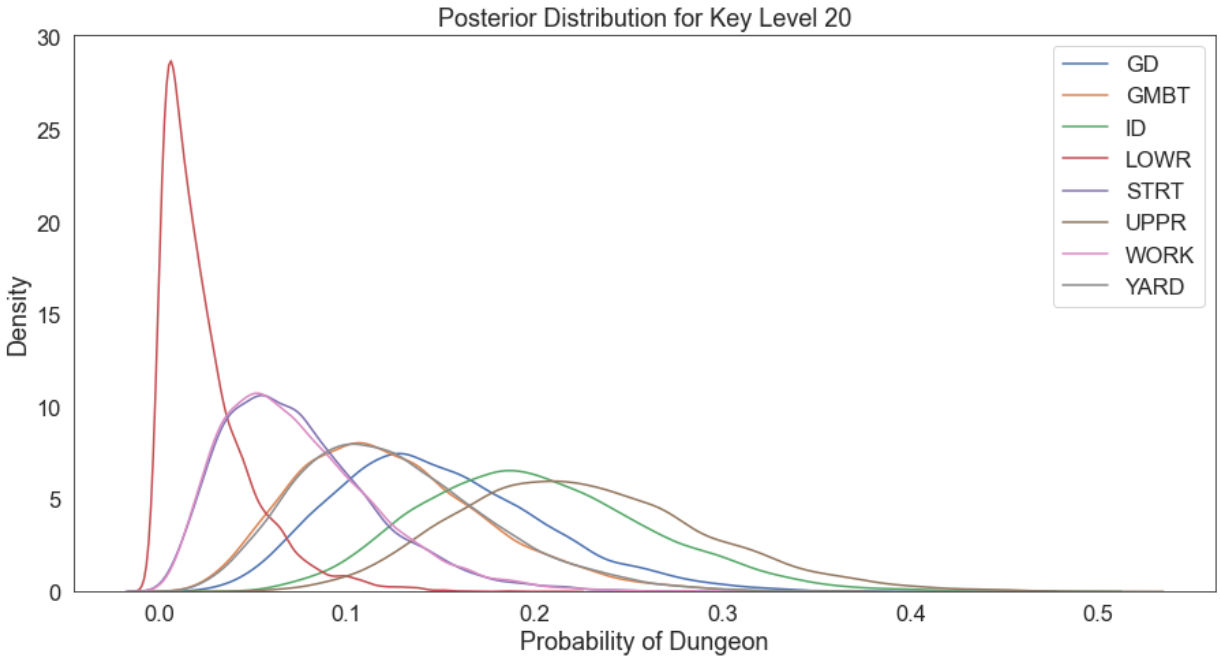

Figure 37: Posterior distributions for the probabilities of each dungeon at the +20 level.

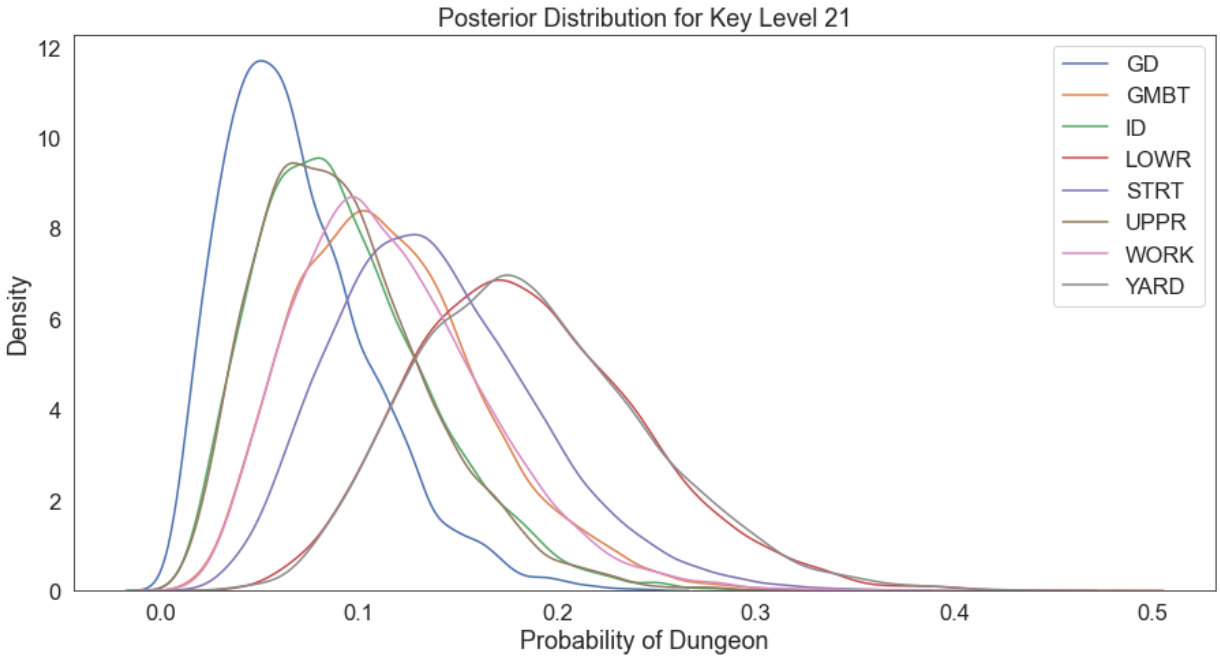

Figure 38: Posterior distributions for the probabilities of each dungeon at the +21 level.

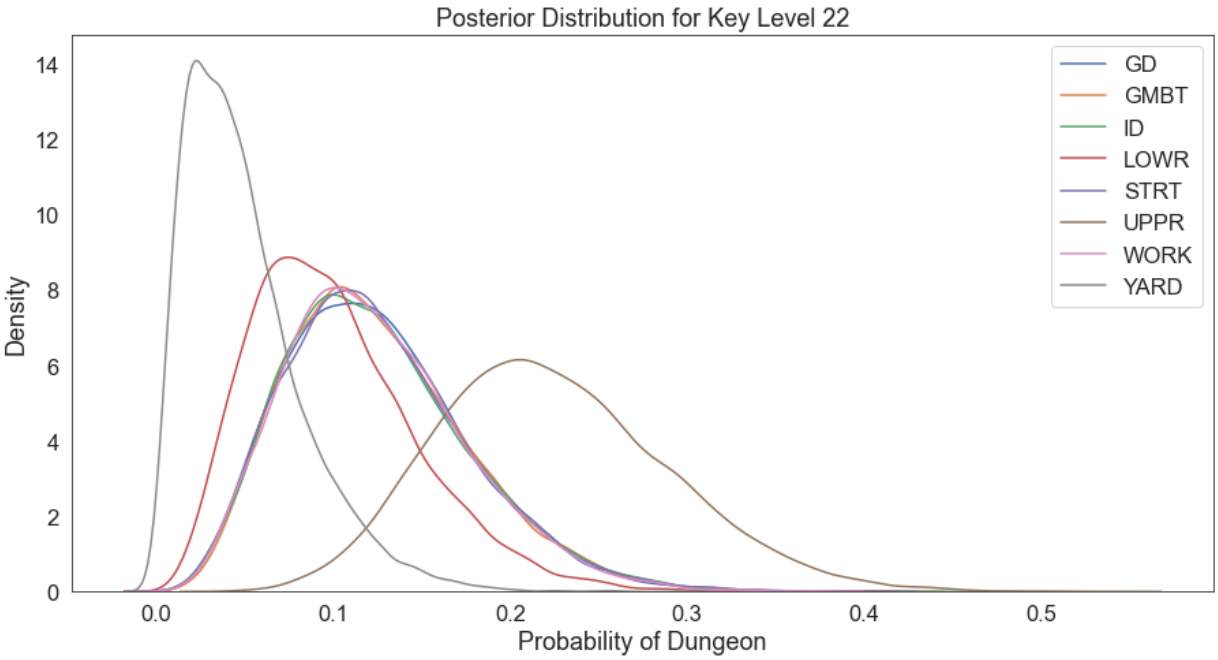

Figure 39: Posterior distributions for the probabilities of each dungeon at the +22 level.

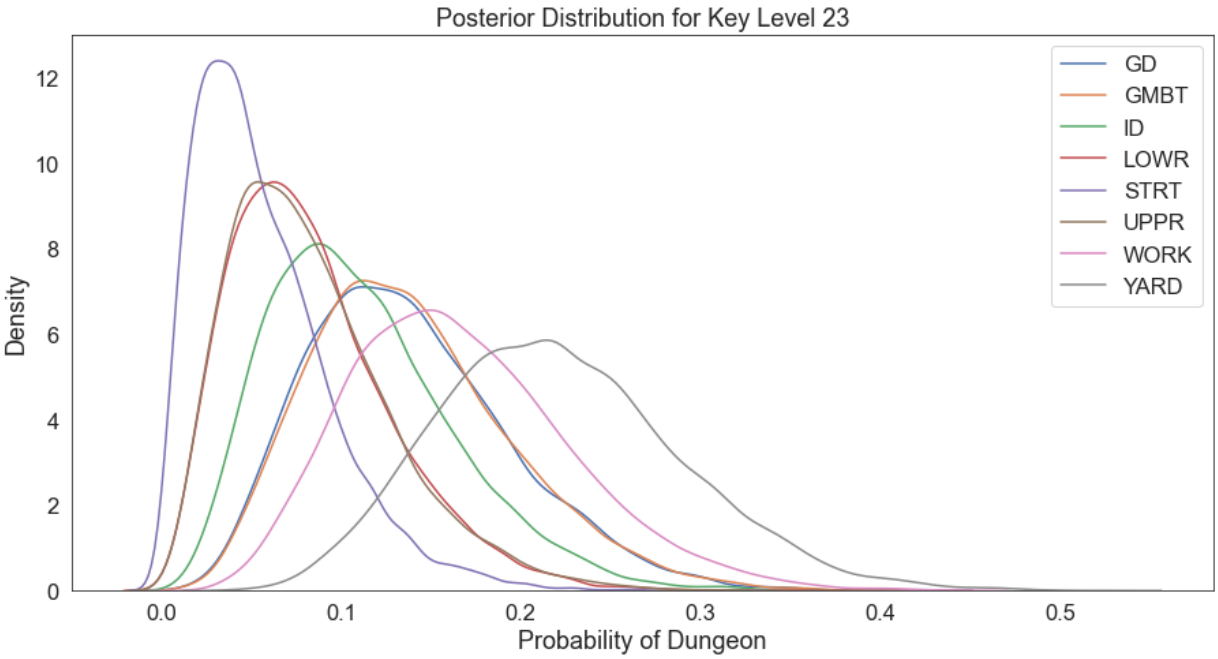

Figure 40: Posterior distributions for the probabilities of each dungeon at the +23 level.

5 Analysis

The analysis of the question "Are the categories (dungeons) evenly distributed?" was done in three different ways:

1. Comparing credible intervals for each dungeon probability parameter p_i
2. The Bayes factor
3. A χ^2 test for comparison between Bayesian and frequentist points of view

Multiple methods were done as way to compare potentially different outcomes depending on the analysis method chosen. In addition, if it can be determined that the categorical probabilities are *not* uniformly distributed, do the distributions vary on a week by week basis, or do the distributions follow the same shape from one week to the next?

5.1 Credible Intervals

For this analysis, the central 95% credible intervals were calculated using the posterior distribution of the model. The 10,000 samples of the probability vectors used to create the posterior distribution plots were used for these calculations. The credible intervals can be seen in Tables (1)-(4).

Subset of Data	Dungeon			
	GD	GMBT	ID	LOWR
All Data	(0.1021, 0.1657)	(0.1106, 0.1753)	(0.1038, 0.1671)	(0.0971, 0.1605)

Subset of Data	Dungeon			
	STRT	UPPR	WORK	YARD
All Data	(0.0693, 0.1246)	(0.0954, 0.1588)	(0.1003, 0.1633)	(0.0879, 0.1470)

Table 1: 95% credible intervals for all data.

When looking at specific affixes, no credible intervals are completely disjoint, and as such cannot be used to declare a significant deviation from uniform categorical probabilities. This is also true when looking at the credible intervals for all data.

When doing the analysis using the 95% credible intervals, in general the data does not seem to suggest a deviation from uniformly distributed categorical probabilities. However, Tables (2) and (4) have values which can be seen as significant deviations from this scheme. The intervals are highlighted such that the interval highlighted in green has a 95% credible interval that is strictly less than the 95% credible intervals highlighted in blue within the same row. As such, the STRT dungeon within the "Fortified, Spiteful, Necrotic" has no overlap and is strictly lower than GD, ID, LOWR, and YARD. This would suggest that during this affix set, there the probability of getting specific keystones is not uniformly distributed. Similarly, during the "Tyrannical, Bolstering, Explosive" affix set, GD has a credible interval that is strictly less than, with no overlap, the credible interval of LOWR. While this could indicate a deviation from uniform categorical probabilities, it is not as decisive as the "Fortified, Spiteful, Necrotic" affix set.

Subset of Data	Dungeon			
	GD	GMBT	ID	LOWR
Fortified, Spiteful, Necrotic	(0.0967, 0.2538)	(0.0437, 0.1678)	(0.0877, 0.2403)	(0.0956, 0.2550)
Tyrannical, Inspiring, Quaking	(0.1151, 0.2846)	(0.0606, 0.1982)	(0.0434, 0.1692)	(0.0203, 0.1205)
Fortified, Sanguine, Grievous	(0.0909, 0.2430)	(0.1297, 0.2998)	(0.1103, 0.2704)	(0.0254, 0.1321)
Tyrannical, Bolstering, Explosive	(0.0088, 0.1027)	(0.0529, 0.2010)	(0.0838, 0.2545)	(0.1040, 0.2875)
Fortified, Bursting, Storming	(0.0262, 0.1545)	(0.0261, 0.1537)	(0.0664, 0.2362)	(0.0790, 0.2555)
Tyrannical, Raging, Volcanic	(0.0575, 0.1921)	(0.1200, 0.2876)	(0.0262, 0.1331)	(0.0755, 0.2210)

Subset of Data	Dungeon			
	STRT	UPPR	WORK	YARD
Fortified, Spiteful, Necrotic	(0.0076, 0.0859)	(0.0204, 0.1197)	(0.0609, 0.1986)	(0.1156, 0.2831)
Tyrannical, Inspiring, Quaking	(0.0601, 0.1970)	(0.0866, 0.2416)	(0.0864, 0.2420)	(0.0446, 0.1680)
Fortified, Sanguine, Grievous	(0.0492, 0.1740)	(0.0487, 0.1747)	(0.0403, 0.1608)	(0.0324, 0.1460)
Tyrannical, Bolstering, Explosive	(0.0245, 0.1424)	(0.0836, 0.2577)	(0.0841, 0.2541)	(0.0428, 0.1833)
Fortified, Bursting, Storming	(0.0687, 0.2340)	(0.0789, 0.2514)	(0.0566, 0.2164)	(0.0574, 0.2139)
Tyrannical, Raging, Volcanic	(0.0576, 0.1915)	(0.0658, 0.2069)	(0.0664, 0.2061)	(0.0491, 0.1766)

Table 2: 95% credible intervals for observed weekly affix sets.

Subset of Data	Dungeon			
	GD	GMBT	ID	LOWR
Fortified	(0.0989, 0.1916)	(0.0915, 0.1809)	(0.1186, 0.2185)	(0.0876, 0.1750)
Tyrannical	(0.0821, 0.1688)	(0.1050, 0.1990)	(0.0664, 0.1466)	(0.0852, 0.1724)

Subset of Data	Dungeon			
	STRT	UPPR	WORK	YARD
Fortified	(0.0537, 0.1280)	(0.0648, 0.1438)	(0.0729, 0.1542)	(0.0910, 0.1820)
Tyrannical	(0.0670, 0.1474)	(0.1050, 0.1976)	(0.1052, 0.1994)	(0.0631, 0.1404)

Table 3: 95% credible intervals for observed individual affixes.

Similarly with affix sets, there is also an instance of deviation within the key level credible intervals. At the +20 key level, LOWR has a credible interval that is strictly less than both ID and UPPR. Hence, there is a reasonable chance that the categorical probabilities are not uniformly distributed at the +20 key level.

5.2 Bayes Factor

The Bayes factor, K, is defined as the ratio of the marginal likelihoods of two competing models, M_1 and M_2. This value can then be used as a comparison of the effectiveness of the models. It is defined as

$$K = \frac{\Pr(\text{Data}|M_2)}{\Pr(\text{Data}|M_1)}$$

Here, Model 1 can be set to be the likelihood under a null hypothesis, so that a Bayesian hypothesis test can be performed. Model 2 would then be the model defined in Section (3).

Now, since the prior assumption in the model of Section (3) was that the categorical probabilities were all equivalent, this becomes the null hypothesis, i.e., $H_0 : p_1 = p_2 =$

Subset of Data	Dungeon			
	GD	GMBT	ID	LOWR
+2	(0.0895, 0.2309)	(0.0815, 0.2169)	(0.0579, 0.1811)	(0.0882, 0.2307)
+15	(0.0704, 0.3690)	(0.0712, 0.3746)	(0.0011, 0.1438)	(0.0259, 0.2720)
+16	(0.0541, 0.2610)	(0.0158, 0.1691)	(0.0868, 0.3192)	(0.0887, 0.3198)
+17	(0.0444, 0.2472)	(0.0448, 0.2475)	(0.1131, 0.3724)	(0.0602, 0.2785)
+18	(0.0363, 0.2534)	(0.0354, 0.2590)	(0.0356, 0.2580)	(0.0375, 0.2589)
+19	(0.0395, 0.2713)	(0.0796, 0.3594)	(0.0088, 0.1771)	(0.0084, 0.1741)
+20	(0.0589, 0.2743)	(0.0420, 0.2401)	(0.0944, 0.3330)	(0.0007, 0.0912)
+21	(0.0150, 0.1607)	(0.0400, 0.2269)	(0.0265, 0.1948)	(0.0853, 0.3112)
+22	(0.0420, 0.2391)	(0.0440, 0.2423)	(0.0413, 0.2433)	(0.0292, 0.2101)
+23	(0.0477, 0.2653)	(0.0481, 0.2670)	(0.0325, 0.2309)	(0.0180, 0.1904)

Subset of Data	Dungeon			
	STRT	UPPR	WORK	YARD
+2	(0.0592, 0.1821)	(0.0661, 0.1947)	(0.0508, 0.1686)	(0.0508, 0.1689)
+15	(0.0283, 0.2758)	(0.0264, 0.2708)	(0.0099, 0.2156)	(0.0265, 0.2699)
+16	(0.0058, 0.1277)	(0.0155, 0.1648)	(0.0716, 0.2911)	(0.0425, 0.2345)
+17	(0.0063, 0.1368)	(0.0171, 0.1776)	(0.0447, 0.2476)	(0.0308, 0.2135)
+18	(0.0077, 0.1690)	(0.0351, 0.2564)	(0.1194, 0.4145)	(0.0076, 0.1677)
+19	(0.0222, 0.2297)	(0.0390, 0.2731)	(0.0802, 0.3587)	(0.0218, 0.2293)
+20	(0.0159, 0.1719)	(0.1109, 0.3677)	(0.0166, 0.1739)	(0.0439, 0.2448)
+21	(0.0541, 0.2543)	(0.0268, 0.1940)	(0.0409, 0.2271)	(0.0868, 0.3131)
+22	(0.0417, 0.2444)	(0.1112, 0.3644)	(0.0426, 0.2388)	(0.0064, 0.1328)
+23	(0.0071, 0.1510)	(0.0182, 0.1934)	(0.0658, 0.3047)	(0.1031, 0.3658)

Table 4: 95% credible intervals for various key levels.

$\cdots = p_k$. Under this null hypothesis, the marginal likelihood becomes

$$\Pr(\text{Data}|M_1) = \text{Multinomial}(\mathbf{Y}|\mathbf{p_0})$$

where $p_{01}, \ldots, p_{0k} = \frac{1}{k}$, where in this case $k = 8$ and $\mathbf{Y}$ is also as defined in Section (3).

The likelihood for the data is calculated as

$$\Pr(\text{Data}|M_2) = \text{Multinomial}(\mathbf{Y}|\mathbf{p})$$

However, this likelihood is over an unknown parameter, and hence has a prior distribution. Thus to get the marginal likelihood, the likelihood needs integrated over all possible prior distributions.

$$\Pr(\text{Data}|M_2) = \int \text{Multinomial}(\mathbf{Y}|\mathbf{p}) \times \text{Dirichlet}(\mathbf{p}|\boldsymbol{\alpha})\, d\mathbf{p}$$

$$\Pr(\text{Data}|M_2) = \int \binom{n}{Y_1, \ldots, Y_k} \left(\prod_{j=1}^{k} p_j^{Y_j} \right) \left(\frac{1}{\text{B}(\boldsymbol{\alpha})} \prod_{j=1}^{k} p_j^{\alpha_j - 1} \right) d\mathbf{p}$$

$$\Pr(\text{Data}|M_2) = \binom{n}{Y_1, \ldots, Y_k} \frac{\mathrm{B}(\boldsymbol{\alpha'})}{\mathrm{B}(\boldsymbol{\alpha})} \int \frac{1}{\mathrm{B}(\boldsymbol{\alpha'})} \prod_{j=1}^{k} p_j^{\alpha_j + Y_j - 1} \, d\mathbf{p}$$

$$\Pr(\text{Data}|M_2) = \binom{n}{Y_1, \ldots, Y_k} \frac{\mathrm{B}(\boldsymbol{\alpha'})}{\mathrm{B}(\boldsymbol{\alpha})}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha'}$ are as defined in Section (3) and

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}$$

Thus, the Bayes factor comes out to be

$$K = \frac{\binom{n}{Y_1, \ldots, Y_k} \frac{\mathrm{B}(\boldsymbol{\alpha'})}{\mathrm{B}(\boldsymbol{\alpha})}}{\binom{n}{Y_1, \ldots, Y_k} \prod_{j=1}^{k} (\frac{1}{k})^{Y_j}}$$

$$K = \frac{\mathrm{B}(\boldsymbol{\alpha'})}{\mathrm{B}(\boldsymbol{\alpha})} k^{\sum_{j=1}^{k} Y_j}$$

$$K = \frac{\mathrm{B}(\boldsymbol{\alpha'})}{\mathrm{B}(\boldsymbol{\alpha})} k^{n}$$

The Bayes factor for all of the data was calculated to be $K = 6.178 \times 10^{-6}$. The other Bayes factors can be seen in Tables (5)-(7).

Affix Set	Bayes Factor
Fortified, Spiteful, Necrotic	1.219×10^{0}
Tyrannical, Inspiring, Quaking	1.377×10^{-2}
Fortified, Sanguine, Grievous	9.420×10^{-2}
Tyrannical, Bolstering, Explosive	9.933×10^{-2}
Fortified, Bursting, Storming	2.979×10^{-3}
Tyrannical, Raging, Volcanic	3.461×10^{-3}

Table 5: Bayes factor for observed weekly affix sets.

Affix	Bayes Factor
Fortified	2.423×10^{-4}
Tyrannical	9.411×10^{-5}

Table 6: Bayes factor for individual affixes.

To begin analyzing the resulting Bayes factors, a measure of significance is needed. Here, the scale mentioned in Kass and Raftery, 1995 [1] will be used. For convenience, it can be found in Table (8). B_{10} in the table is equivalent to K in the above calculations.

With these interpretations of the Bayes factor, there are few points where there can be any evidence against the null hypothesis. These points are the affix set "Fortified, Spiteful, Necrotic" and at the +20 keystone level. Even still, these two points fall under the "Not worth more than a bare mention" category.

Key Level	Bayes Factor
+2	3.256×10^{-4}
+15	1.686×10^{-1}
+16	1.610×10^{-1}
+17	7.660×10^{-2}
+18	1.134×10^{-1}
+19	8.722×10^{-2}
+20	1.100×10^{0}
+21	1.839×10^{-2}
+22	3.771×10^{-2}
+23	7.888×10^{-2}

Table 7: Bayes factor for various key levels.

$\log_{10}(B_{10})$	B_{10}	Evidence Against H_0
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
$\geq$2	$\geq$100	Decisive

Table 8: Interpretation of Bayes factor as described in Kass and Raftery, 1995 [1].

5.3 χ^2-Test

In frequentist statistics, a χ^2-test can be performed on categorical data to check the distribution against a null hypothesis. Here, the null hypothesis stays the same in that $H_0 : p_1 = p_2 = \cdots = p_k$. The test is performed by taking the sum of squared differences between the observed counts, n_i and the expected observed counts $E(n_i)$ under the null hypothesis, taking the ratio of this sum to the expected counts, and then comparing to a χ^2 distribution with $k-1$ degrees of freedom, where k is the number of categories.

$$\sum_{i=1}^{k} \frac{(n_i - E(n_i))^2}{E(n_i)} \sim \chi^2_{k-1}$$

The p-value when using the entire dataset was calculated to be $p = 0.6623$. The other p-values for the same subsets of data as before can be seen in Tables (9)-(11).

Set of Affixes	p-Value
Fortified, Spiteful, Necrotic	0.0216
Tyrannical, Inspiring, Quaking	0.2414
Fortified, Sanguine, Grievous	0.0467
Tyrannical, Bolstering, Explosive	0.1386
Fortified, Bursting, Storming	0.7798
Tyrannical, Raging, Volcanic	0.4369

Table 9: p-values for observed weekly affix sets.

Affix	p-Value
Fortified	0.3549
Tyrannical	0.5354

Table 10: p-values for individual affixes.

Key Level	p-Value
+2	0.9029
+15	0.5329
+16	0.2322
+17	0.3354
+18	0.3326
+19	0.4980
+20	0.0853
+21	0.6044
+22	0.4827
+23	0.3857

Table 11: p-values for various keystone levels.

P-values are, of course, up for interpretation. Here, any p-value less than 0.05 will be considered significant. In this scenario, there two subsets of data that have a significant deviation from the null hypothesis. These sets are the "Fortified, Spiteful, Necrotic" affix set and the "Fortified, Sanguine, Grievous" affix set. All other subsets of the data have insignificant p-values.

5.4 Method Outcomes and Comparisons

When using 95% credible intervals, the null hypothesis that the categorical probabilities are uniform could be rejected in the cases the "Fortified, Spiteful, Necrotic" affix set, the "Tyrannical, Bolstering, Explosive" affix set, and the +20 keystone level. For the Bayes factor, no subset had a Bayes factor high enough to be considered worth more than a bare mention, and as such the null hypothesis could not reasonably be rejected in any case. For the frequentist comparison, the null hypothesis could be rejected for the "Fortified, Spiteful, Necrotic" affix set and the "Fortified, Sanguine, Grievous" affix set.

An interesting observation is the difference in outcomes based on the methodology used. a frequentist approach would decisively reject the null hypothesis in both of its scenarios because of the p-value being less than the critical value of 0.05. However, the Bayesian approach of Bayes factors says that there is not enough evidence in any scenario to reject the null hypothesis, unless a very lenient mindset is taken. The case of credible intervals is more of an interesting way to look at the data, however doesn't work well for actually determining significant differences in this situation because of the interdependence on the different p_i, since there is the constraint that $\sum_{i=1}^{k} p_i = 1$. Hence, if a value of p_i is on the lower end, then by necessity, the other values must tend towards their higher ends of their intervals.

When considering all frames of mind, the only point that one can reasonably reject the null hypothesis is in the case of the "Fortified, Spiteful, Necrotic" affix set. Even still, it seems unlikely that 5 of the 6 affix sets would be uniformly distributed and just one would not be. This particular affix set is considered difficult, however, there is no evidence that because of this certain more difficult dungeons were given less frequently as compensation. Hence, it is likely that this is just an outlier situation.

Due to there not being decisive evidence of non-uniformity within weeks, the analysis of differences between weeks was left not done. If all situations, e.g., weekly affix sets, are deemed uniform, then by necessity the situations cannot be deemed different from each other.

6 Further Study

The largest point of further study would come with more data. With having to fit data into eight categories, and then further subsetting the data by conditioning on the keystone level, individual affixes, or affix set, the data set sizes used for analysis become even smaller to the point that the within category variation of the posterior distributions becomes large. More data collection would also allow for every affix to be conditioned on rather than just two of them, as when conditioning on individual affixes, this results in identical distributions to affix sets since most affixes only had the chance to appear once during the data collection cycle.

Another point of further study would be expanding the scope of data collection. For this analysis, only the first keystone a player received each week was collected. However, the data collection could be expanded to allow for subsequent keystones by recording the keystone the player received, as well as what the keystone was beforehand. The likelihood could then be generated as a sum of multinomial random variables, the non-first keys having $k - 1$ categories. This would allow for a larger sample size to be constructed, since players would not be limited to contributing one data point per week.

References

[1] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.