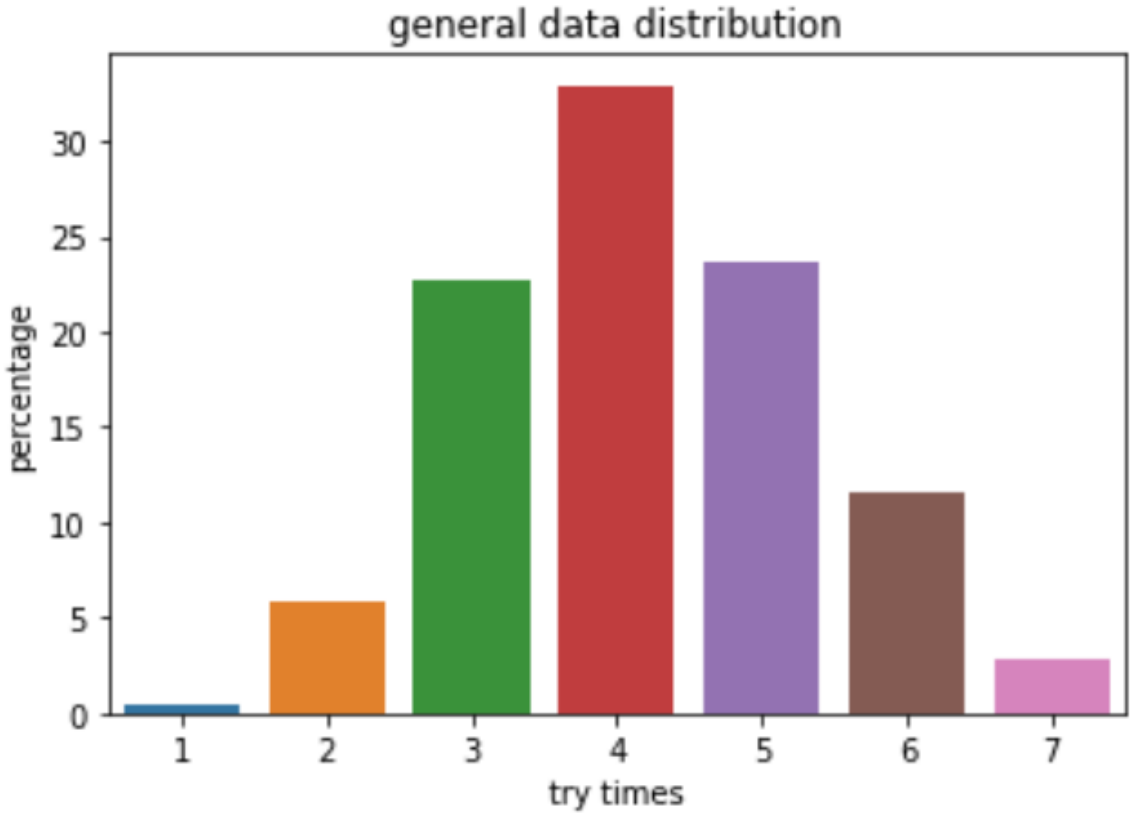
思路：

1. 题目的数据集（try time）分析，推断在推特上分享成绩的人只是少量的成功的人
2. 建立算法模型分析 得到单词期望的预测走向图
3. 验证模型：相关系数计算和假设检验
4. 进一步分析：直方图的绘制（大山和小山的差别）：冰山一角
5. 推断心理：什么样的人群喜欢分享成绩

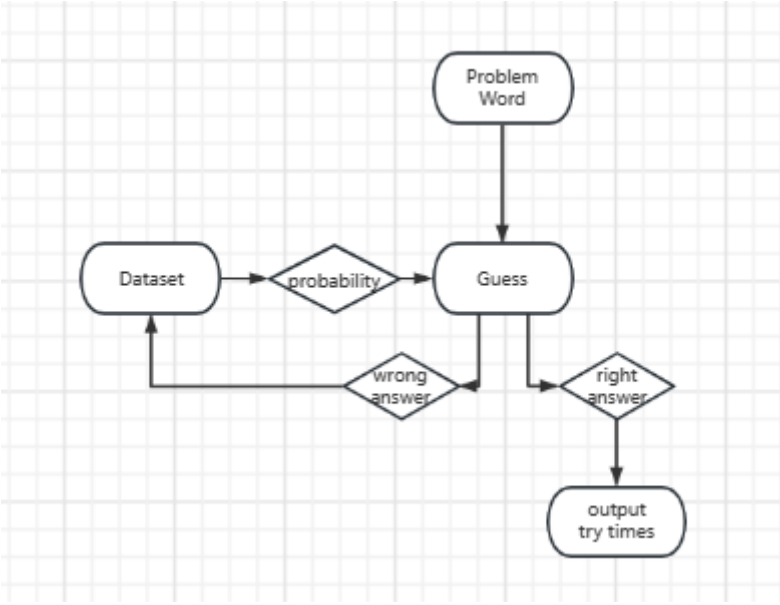
4 有趣的事情

由于题目的原始数据是从推特上直接得出的，来源于选择在推特分享自己成绩的用户，因此，题目数据并没有覆盖游戏的全部玩家。游戏一共有六次机会，在六次中猜到正确答案的玩家即视为通关。而对于题目信息的总体数据，我们可以得出大部分分享数据的玩家都能够在六次机会中通关（97.2%），而这不符合一般游戏的难度设定规则。**因此，我们假设：在大多数情况下，在推特上分享出自己的成绩的用户大部分为成绩较好的用户，而没通关或成绩较差的用户则不倾向于在推特上分享自己的游戏数据。**

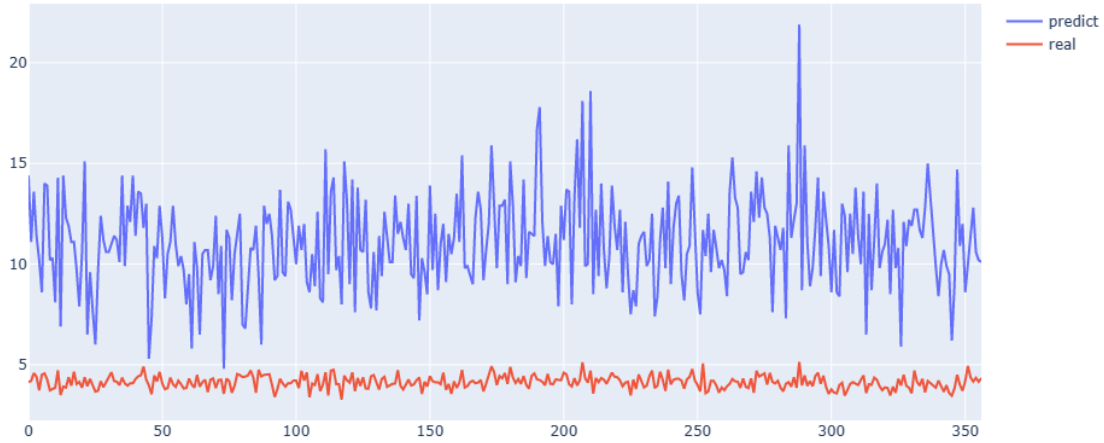


为了验证我们的假设，我们利用计算机模仿人类在游戏过程中的选择于判断。来得出在一般情况下大家玩此游戏的通过率情况以及得分的分布。**从一般人的思路出发，我们假设人在选择填入的单词满足词频的大小分布。**

因此，我们利用wordle的单词库，以及在wolftram Datafrequency上获取的英语单词词频数据集，做出算法来模拟人类在决策中的思路：具体的流程图如下图所示：



我们对每个单词运算20次，求出平均期望，与题目已知单词的期望的做了比较，并作出相关性分析与假设检验，如下图所示：



	test	predict
test	1.000000	0.530377
predict	0.530377	1.000000

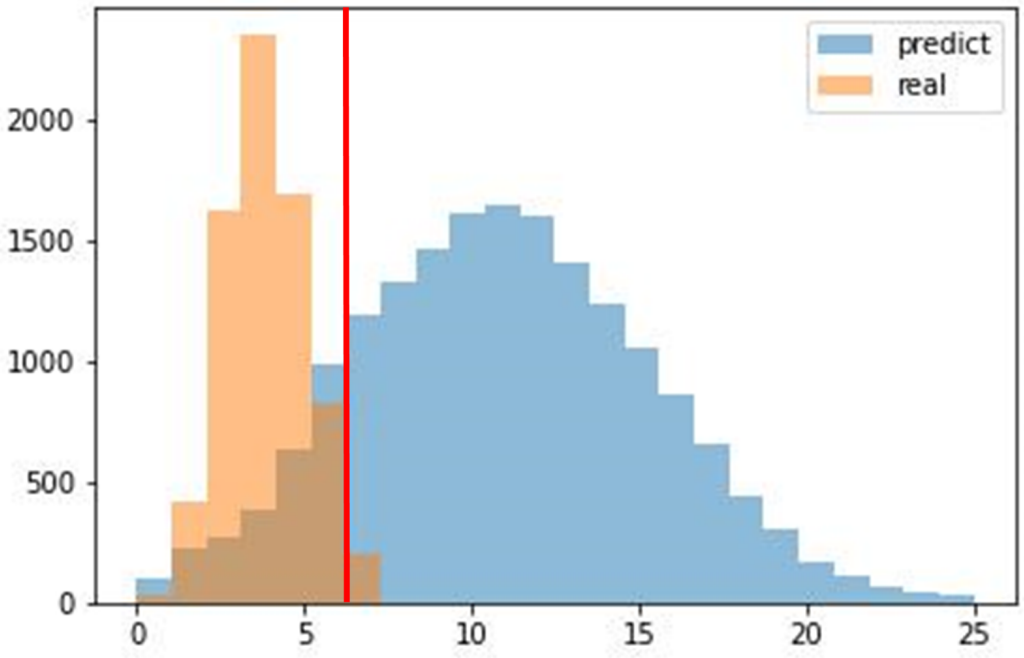
相关性

		test	predict
test	皮尔逊相关性	1	.530**
	Sig.（双尾）		.000
	平方和与叉积	39.038	129.098
	协方差	.110	.363
	个案数	357	357
predict	皮尔逊相关性	.530**	1
	Sig.（双尾）	.000	
	平方和与叉积	129.098	1517.698
	协方差	.363	4.263
	个案数	357	357

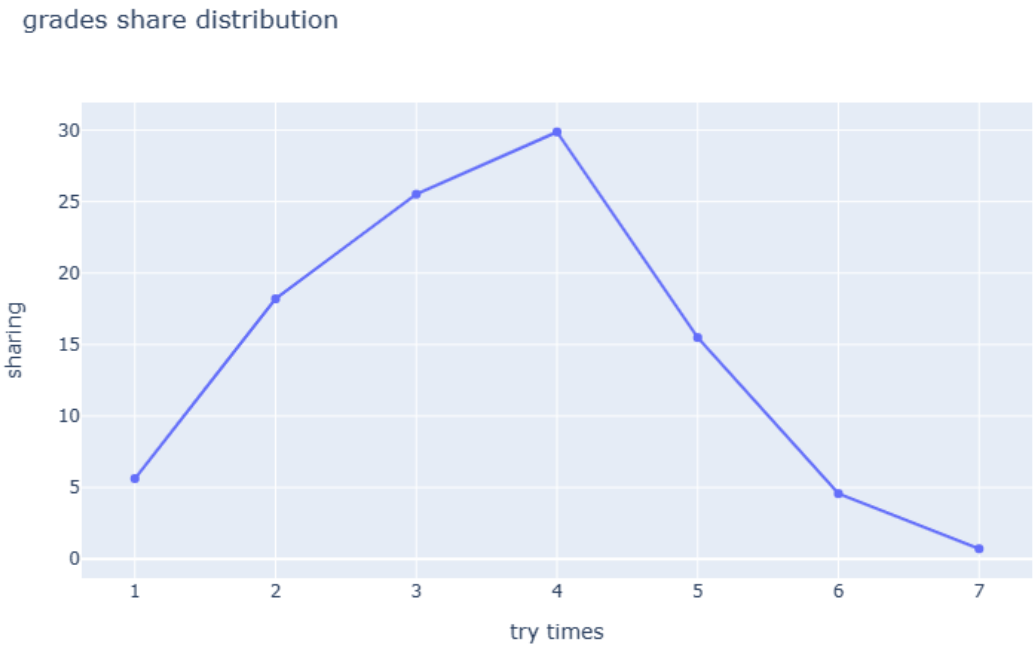
**. 在 0.01 级别（双尾），相关性显著。

在预测折线图中，预测值的期望显著地大于真实值的期望，差值为 11.04750700280112 - 4.121081322566646，其中，两组数据各单词的期望拟合度达到0.53，由皮尔逊相关系数检验得出，p=0.000<0.01,在99%的概率上认为两者呈现显著的相关性。模型拟合程度较高，假设具有一定程度的正确性

进一步地，依据此模型，我们对每个单词的次数散布区间做出预测，将数据汇总后总体数据的直方图如下图所示：由图像可得，分享数据的用户只是全体数据的冰山一角，在显示的数据集后，存在着大量失败、未成功猜出单词的用户。在推特上分享自己数据的用户大部分具有游戏成绩且游戏成绩较好



在该模型的基础上，我们进一步对用户在什么情况下更喜欢在推特上分享成绩进行了调研，由 $\frac{\text{真实值}}{\text{预测值}}$ 并对数据做归一化处理，得到折线图如下所示：



在误差允许范围内，我们发现：与期望结果不同，分享数据的比例并没有呈现单调递减的趋势，而是先增后减，并在x=4处达到峰值。虽然整体上成绩较好的玩家更喜欢发表数据，但当成绩过好（尝试次数靠近1）时，比例反而下降。我们推断：此类群体猜出单词更大程度上依据于概率，而非题目的信息量。迅速的成功并不能让用户体会到游戏的乐趣，而3~5区间段的用户猜出数据更多依靠思路与题目不断更新的信息量，在该区间猜出单词对用户所带来的成就感越强，符合心理学的推断与预期