量化梯度 v2

slcju

2022.6

1 上一版简述及其存在的问题

一个英雄的强度由胜率与 bp 率联合决定，二者应为乘积的关系。下文中 w 指胜率，p 指出场率，b 指禁用率，单位均为百分点。根据这三个量，设计一个指标 S 对英雄[1]进行排名。

首先根据非 ban 必选的含义，将 p 和 b 合为一个称为 bp 率的指标 P。[2]

$$P_0(p,b) = \frac{p}{100-b} \tag{1}$$

$$P = \frac{P_0}{\mu_{P_0}} \tag{2}$$

然后用函数将胜率和 bp 率的实际意义传达到总分 S 中。

$$S = f_w(w_t) \cdot f_P(P_t) \tag{3}$$

注意，这里使用了 w_t 和 P_t 而不是 w 和 P。下标 t 表示标准化的值。**标准化用于解决两个指标尺度不一致的问题**，这是上一版没有深入考虑的。

分析了胜率的实际意义，选用 softplus 函数作为 f_w：

$$f_w(w_t) = \log_2(1 + 2^{w_t}) \tag{4}$$

对于 bp 率，上一版相当于直接令 $f_P(P_t) = P$，引入标准化后，这应为

$$f_P(P_t) = P_t \tag{5}$$

2 数据标准化

一个常用的方式是，将数据线性映射到标准正态分布，即减去均值，再除以标准差[3]：

$$w_t = \frac{w - \mu_w}{\sigma_w} \tag{6}$$

[1] 或（英雄，分路，技能）组合，后文一般简称英雄。

[2] 上一版中 μ_{P_0} 被记为 C。这里将 P_0/μ_{P_0}（即上一版中的 P/C）表示为一个整体 P，是为了处理和表达上的方便。在梯度坐标图展示时，展示 P_0 而不是 P，以尽可能让坐标图浅显易懂。

[3] Z-score 标准化。标准差也可叫波动率。上一版相当于直接认定 $\mu_w = 50, \sigma_w = 1$。

$$P_t = \frac{P - \mu_P}{\sigma_P} \tag{7}$$

这种方法的前提是，原始数据应大致符合或有充分的理由假定为正态分布。那么，胜率和 bp 率是否满足条件呢？这需要**对历史数据进行分析**。

正好之前苏苏给了我一份历史数据。截取 2021.1.1 至 2022.4.25 的数据，去除一些异常值[4]，计算 P[5]，对 w 和 P 拟合概率分布如下[6]。图中横轴是 w 或 P 的取值，纵轴是取值的出现频率。

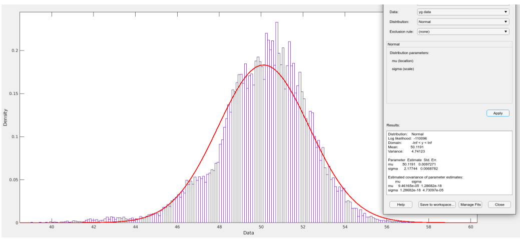

图 1: 1350 分段的胜率分布

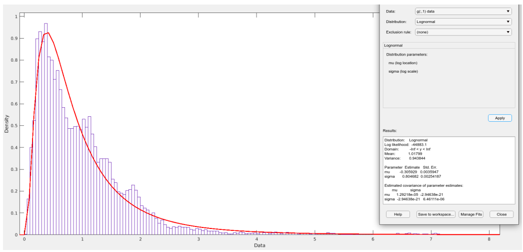

图 2: 1350 分段的 bp 率分布

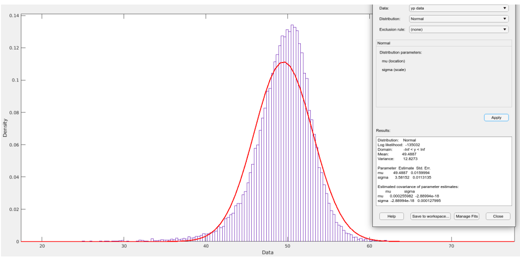

图 3: 顶端局的胜率分布

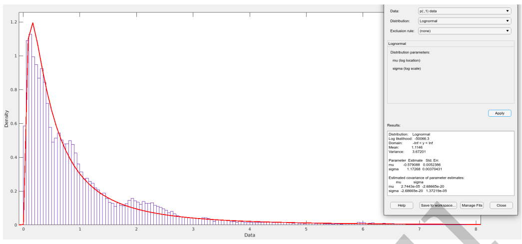

图 4: 顶端局的 bp 率分布

不论是从历史数据还是直觉来看，胜率符合正态分布都是合理的。bp 率则拟合了对数正态分布，即 bp 率取对数后符合正态分布。[7] $\mu_w, \sigma_w, \mu_P, \sigma_P$ 都是拟合可以得到的常数，后文会列出。它们反映了英雄的胜率和 bp 率在不同分段（1350 分段和顶端局）的分布情况。

顶端局的 σ_w 和 σ_P（胜率和 bp 率的波动）都比 1350 分段的大，这可以用顶端局的样本量小来解释。历史数据表明胜率波动大是顶端局环境的固有特征，那么，极高或极低的胜率在顶端局的说服力相比 1350 分段就偏弱。假如马可在顶端局和 1350 分段同样是 45 胜率，那么相比之下在 1350 分段他确实就比较烂了，在顶端局他或许没有那么烂（因为从历史数据来看，在 1350 分段 45 胜率已经是相当靠后了，而在顶端局则并没有那么靠后）。bp 率也有类似的效应。

[4]处理后一共 50110 个数据条目，对应于 480 天，一百零几个英雄。

[5]这里的计算有简化，μ_{P_0} 没有按分路计算，而是以日为单位，计算了每日所有英雄 bp 率的均值。因为这份历史数据并没有细致到分路。这不致有大的误差。

[6]文档中的图都不太清楚，github repo history_analysis 文件夹下有高清大图。这里能看清拟合度较好的概率分布就行。

[7]bp 率取值为 1 表示某英雄的 bp 率等于该日全英雄 bp 率的均值，然而由图可见，数据分布明显左偏，大部分数据小于 1。这就像随便找二十个人跟四个王者策划一起计算薪水均值，那么大多数人的薪水是低于均值的。在王者里好比，顶端局中路火舞西施弈星玉环四人占去了七成的出场率，还有二十个左右的英雄分这剩下的三成，这二十位都会低于平均。另外，这种状况下，往往应该用中位数而不是平均数，但我考虑后还是使用了平均数，因为我想让 T2 以上筛选出足够强力的英雄（如上面例子中的四位），锁定 T2 的含义为“一般强势英雄”，T3 的含义为“大众英雄”。上一版文档中称“T3 的英雄已落后于中线”，应为“均线”，而均线是显著高于中线的。

3 bp 率的处理

bp 率符合对数正态分布，则式 (7) 应为

$$\ln P_t = \frac{\ln P - \mu_P}{\sigma_P} \tag{8}$$

即

$$P_t = \exp(\frac{\ln P - \mu_P}{\sigma_P}) \tag{9}$$

$$P_t = \exp(\frac{\ln P}{\sigma_P}) \cdot \exp(-\frac{\mu_P}{\sigma_P}) \tag{10}$$

$$P_t = P^{1/\sigma_P} \cdot \exp(-\frac{\mu_P}{\sigma_P}) \tag{11}$$

由式 (3) 和式 (5)，$\exp(-\frac{\mu_P}{\sigma_P})$ 这个常数系数并不会对排行产生影响（所有的英雄得分按比例乘以一个正的常数不影响内部排序），因此将其丢弃。丢弃后，总分 S 中 bp 率这一项就比较简单了：

$$S = f_w(w_t) \cdot P^{1/\sigma_P} \tag{12}$$

相比上一版，多了一个指数，反映该分段下 bp 率的固有波动率（历史数据得出的规律）。**bp 率固有波动越大**（比如苏苏取的顶端局），**在总分中 bp 率的权重就越低**，如同上一部分结尾对胜率的分析。

4 胜率的处理

对于胜率，是否将拟合得到的常数 μ_w, σ_w 直接代入式 (6) 再代入式 (4) 呢？且慢，观察一下胜率和 bp 率的联合分布（图 5 和图 6）。横轴为 bp 率，纵轴为胜率，每一个数据点代表历史数据中某日某英雄的胜率和出场率。

观察到，随着 bp 率的增加，胜率所覆盖的“敞口”有缩小的趋势。画一条竖直线，从左往右扫过去，数据点在竖直线上的分布会越来越窄。这暗示，**在 bp 率比较高的情况下，胜率一般不会偏离 50 太多；bp 率比较低时，胜率更可能出现比较大的偏离**。这也是符合直觉的，一方面，出场率低意味着样本量小，波动自然会比较大；另一方面，顶端局 bp 率低的英雄，或是雅典娜狼狗（绝活哥专属），或是甄姬猴子（太弱而无人问津），基本是这两种情况，胜率两极分化。

这个“敞口”与 σ_w 有直接的关系。由以上观察可知，σ_w 不宜被视为一个常数，而是随 P 增大而减小的函数[8]。

以 0.1bp 率的宽度为窗口，从左向右滑动窗口（将上面零宽度的竖直线拓宽到 0.1 的宽度，如图 7 示意），统计窗口内数据点胜率的标准差，作为窗口中点 bp 率对应的 σ_w，将得到的 (P, σ_w) 绘出并拟合函数，如图 8 和图 9 所示。拟合了 $\sigma_w = aP^b + c$ 的函数，其中 b 接近于 -1，近似于带常数的反比例函数。

[8] 将 P 视为自变量，w 视为因变量比较合乎常理，先 bp 再进入游戏决胜负。因此将 bp 率视为独立因素，先处理了 bp 率。

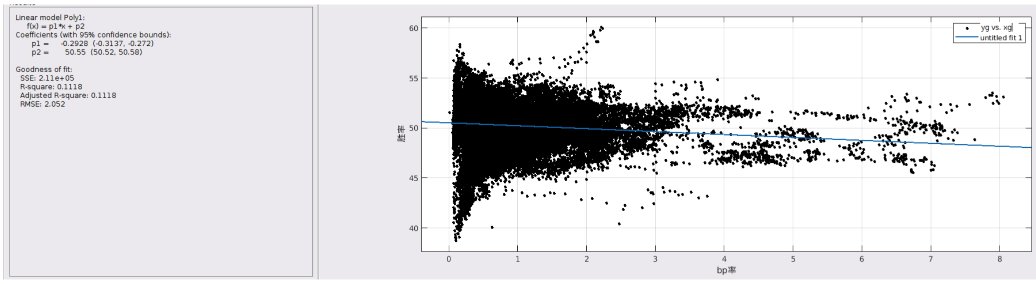

图 5: 1350 分段的胜率-bp 率联合分布

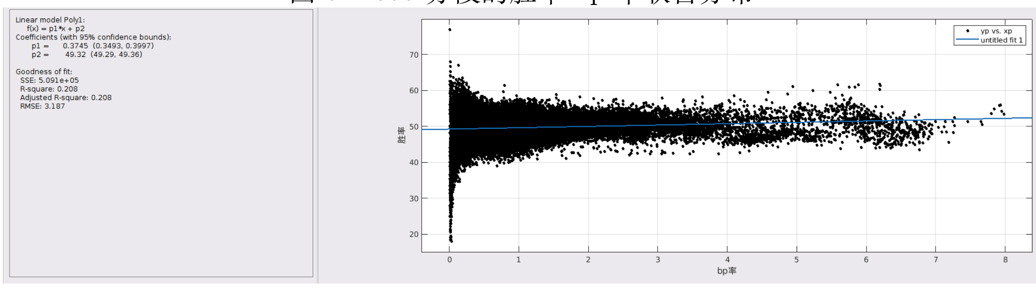

图 6: 顶端局的胜率-bp 率联合分布

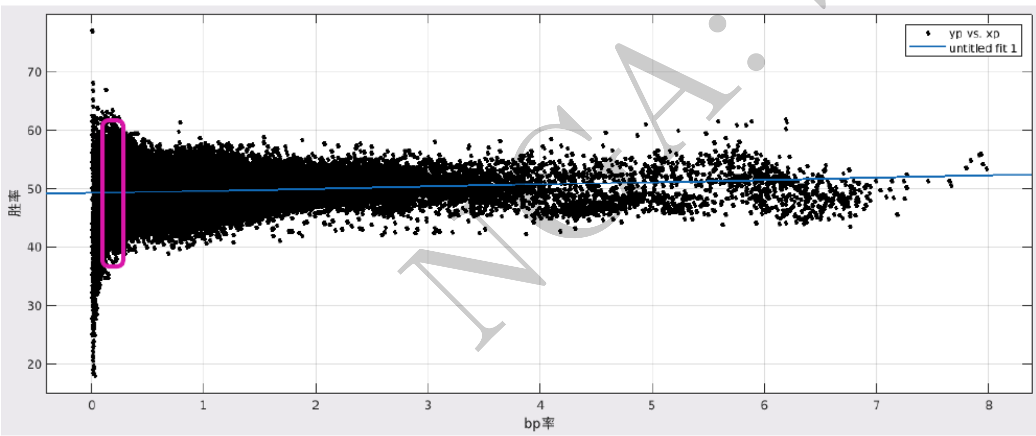

图 7: 胜率波动-bp 率关系的计算

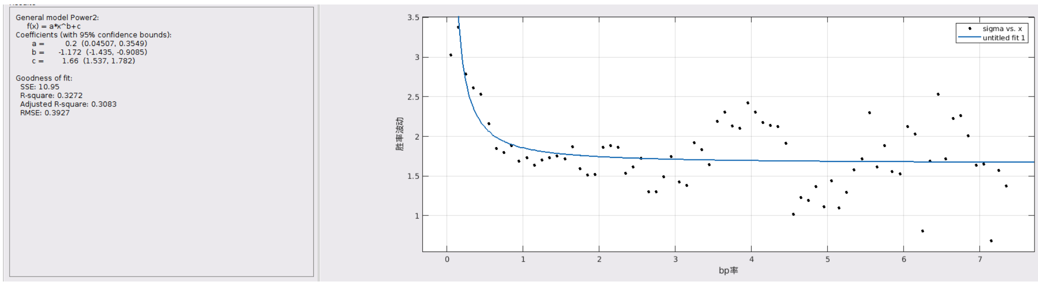

图 8: 1350 分段的胜率波动-bp 率关系

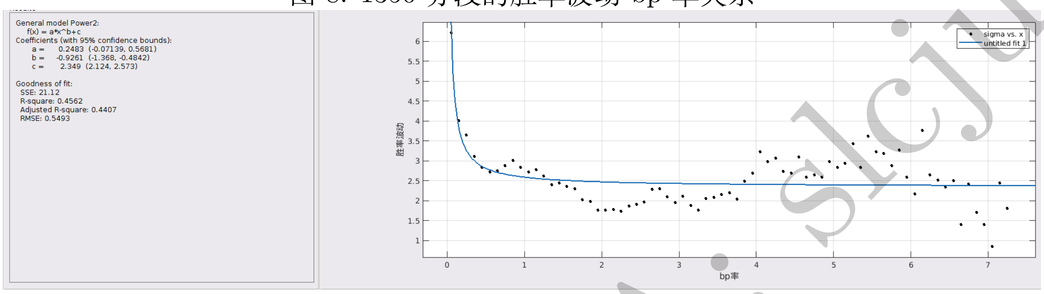

图 9: 顶端局的胜率波动-bp 率关系