

高中学生作业时间与考试成绩的关系

LogCreative*

**Plan Storm Conductor*

2019/12/17

摘要

通过对高二某班级10周时间的作业情况进行统计，得出作业总时间与高考成绩呈正相关，作业偏科程度与高考成绩呈正相关、与期末成绩提升呈二次关系。同班级内学生的每周每科作业时间不符合正态分布。

1 统计过程简述

1.1 背景

笔者发起一项作业监督计划(Plan Storm)，在2016年11月至2017年1月共10周的时间里对高二作业量进行调查。统计学校为安徽省某中学，统计结果可以用来推测普遍情况。

1.2 过程

班委向班级里的每个人发一张待办事项卡(To-do List)，用于记录和监督作业的完成状况，最后用半小时作业时间 T 来表示作业量。而 T 可以根据一个预先设定好评分参考表 [, Appendix A]来统计，例外情况，用 t 表示作业总时长，按下面的公式统计：

$$T = \frac{t}{0.5h} \quad (1)$$

或者用#Paper表示作业页数(A3纸大小)，有下面的经验公式：

$$T = \#Paper \quad (2)$$

1.3 原始数据初筛

通过上述方法，我们就可以得到每个人每周每科的作业量。

原始数据共52人，通过筛选，前9周数据全部完整的最终只有38人。得到的样本总量 $N = 2098$ 。

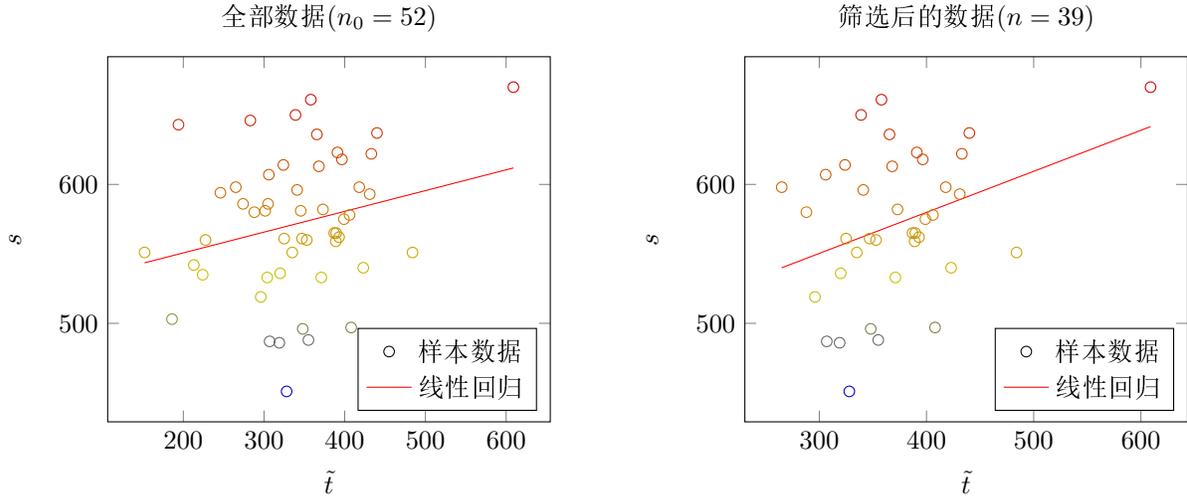
2 作业总量与成绩

2.1 与高考成绩

令9周半小时作业总时间

$$\tilde{T} := \sum_{j=1}^9 T_j \quad (3)$$

图 1: 作业总量与高考成绩散点与线性回归图



通过计算，得出筛选后的数据($n = 39$)的样本作业总时间 \tilde{t} ，高考成绩 s 的均值

$$\begin{aligned}\bar{\tilde{t}} &= \frac{1}{n} \sum_{i=1}^n \tilde{t}_i = 368.577 \\ \bar{s} &= \frac{1}{n} \sum_{i=1}^n s_i = 572.128\end{aligned}$$

并通过下面的步骤计算出样本相关系数：

$$\begin{aligned}S_{\tilde{T}S} &= \sum_{i=1}^n (\tilde{t}_i - \bar{\tilde{t}})(s_i - \bar{s}) = 38895.615 \\ S_{\tilde{T}\tilde{T}} &= \sum_{i=1}^n (\tilde{t}_i - \bar{\tilde{t}})^2 = 156428.769 \\ S_{SS} &= \sum_{i=1}^n (s_i - \bar{s})^2 = 102882.359 \\ \rho_{\tilde{T}S} &= \frac{S_{\tilde{T}S}}{\sqrt{S_{\tilde{T}\tilde{T}}}\sqrt{S_{SS}}} = 0.3066\end{aligned}\quad (4)$$

由此可见， S 与 \tilde{T} 呈正相关。

通过查表 [, Appendix B]可以得到对于显著性水平 $\alpha = 0.10$ 的 F 分布分位点

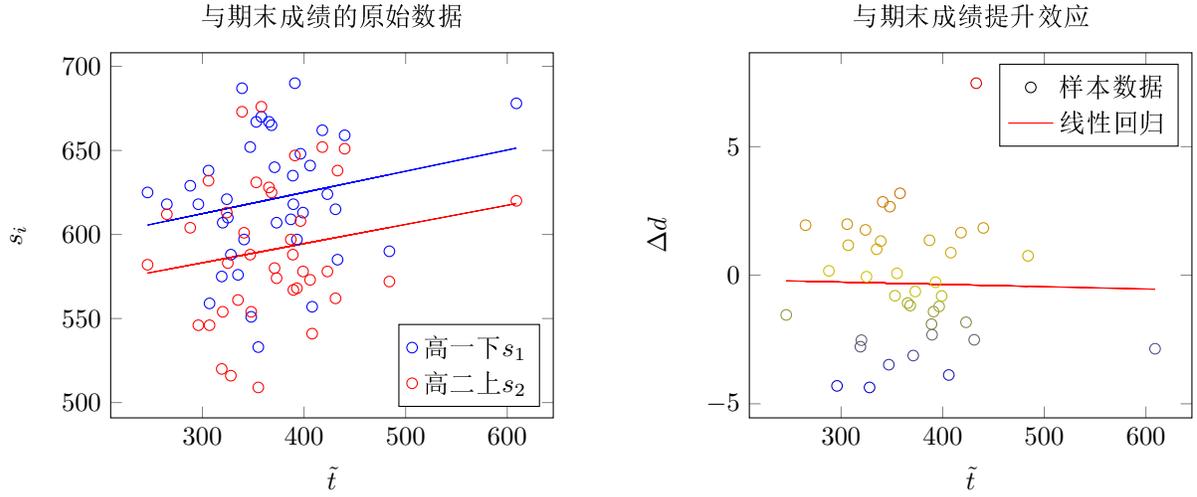
$$F_{0.10}(1, n - 2) = 2.8463\quad (5)$$

以及相关系数临界值

$$r_{0.10}(n - 2) = \sqrt{\frac{1}{\frac{n-2}{F_{0.10}(1, n-2)} + 1}} = 0.2707 < \rho_{\tilde{T}S}\quad (6)$$

所以认为在显著性水平 $\alpha = 0.10$ 下，线性回归显著。

图 2: 作业总量与期末成绩散点与线性回归图



2.2 与期末考试成绩提升

下面讨论作业时间与高一下学期至高二上学期期末考试成绩提升之间的关系。由于两次考试的均值与方差并不相等，下面采用正态分布模型归一化两次考试，以得到客观提升效应值。

通过全年级1300名同学的期末成绩 [, Appendix C], 由Khintchine大数定律, 这些同学的期末成绩独立同分布, 服从大数定律, 故可以通过矩估计法估计 $S_i (i = 1, 2)$ 正态分布的参量 $\hat{\mu}_i, \hat{\sigma}_i^2 (i = 1, 2)$:

$$\begin{aligned}\hat{\mu}_i &= \bar{S}_i \\ \hat{\sigma}_i^2 &= a_2 - \bar{S}_i^2\end{aligned}$$

经过计算得到两次考试的正态分布参量:

$$\begin{aligned}S_1 &\sim N(455.46, 107.85^2) \\ S_2 &\sim N(432.31, 105.59^2)\end{aligned}$$

采用以下公式计算每位学生的成绩偏差值:

$$d_{ij} = \frac{10(s_{ij} - \hat{\mu}_i)}{\hat{\sigma}_i} + 50 \quad (7)$$

并计算两次偏差值的差值, 以定量计算出效应:

$$\Delta d_j = d_{2j} - d_{1j} \quad (8)$$

再对结果进行同样的线性回归处理。得到 \tilde{T} 与期末考试效应 ΔD 的样本相关系数

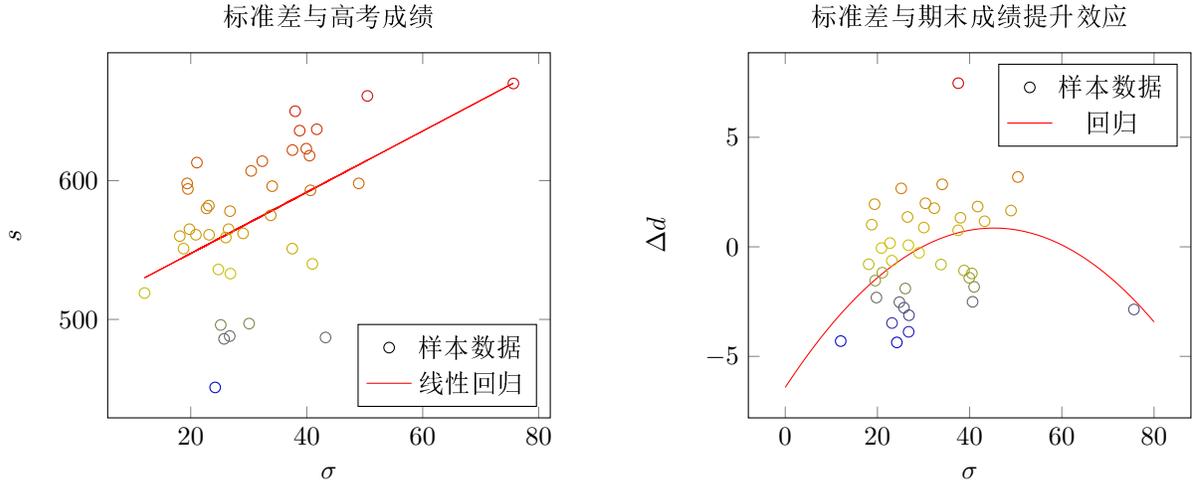
$$\rho_{\tilde{T}\Delta D} = -0.02359 < r_{0.10}(n-2) \quad (9)$$

这个结果表明, ΔD 对 \tilde{T} 负相关, 并且线性回归在显著性水平 $\alpha = 0.10$ 下不显著。并且注意到回归效应极差

$$\widehat{\Delta D}_n = \widehat{\Delta D}_{(n)} - \widehat{\Delta D}_{(1)} = 0.3261 \ll D_n = \Delta D_{(n)} - \Delta D_{(1)} = 11.8323 \quad (10)$$

即回归趋势极差远远小于样本极差, 所以可以认为作业总时间 \tilde{T} 对高一下至高二上的成绩提升没有明显的效应。

图 3: 作业偏科程度与成绩回归图



3 学科分配与成绩

3.1 作业偏科程度与成绩

既然作业总时间与期末成绩提升效应没有明显的关系，那么我们就考察作业偏向性是否与期末成绩提升效应是否有一定的关系。

作业偏向性表征于9周每一科作业时间和的标准差

$$\sigma_j = \sqrt{\frac{1}{6} \sum_{i=1}^6 (\tilde{t}_{ij} - \bar{t}_j)^2} \quad (11)$$

我们将上述标准差与高考成绩的关系采用同样的方法线性拟合，得到相关系数

$$\rho_{\Sigma S} = 0.4979 > r_{0.10}(n-2) \quad (12)$$

说明作业偏科程度与高考成绩在显著性水平 $\alpha = 0.10$ 下显著正相关。

而我们将标准差与期末成绩提升效应 ΔD 采用最小二乘法拟合为二次曲线，有

$$\widehat{\Delta D} = \hat{a} + \hat{b}\Sigma + \hat{c}\Sigma^2 \quad (13)$$

令

$$X = \begin{pmatrix} 1 & \sigma_1 & \sigma_1^2 \\ 1 & \sigma_2 & \sigma_2^2 \\ \vdots & \vdots & \vdots \\ 1 & \sigma_n & \sigma_n^2 \end{pmatrix}, \beta = \begin{pmatrix} a \\ b \\ c \end{pmatrix}, Y = \begin{pmatrix} \Delta d_1 \\ \Delta d_2 \\ \vdots \\ \Delta d_n \end{pmatrix}$$

则

$$Y = X\beta$$

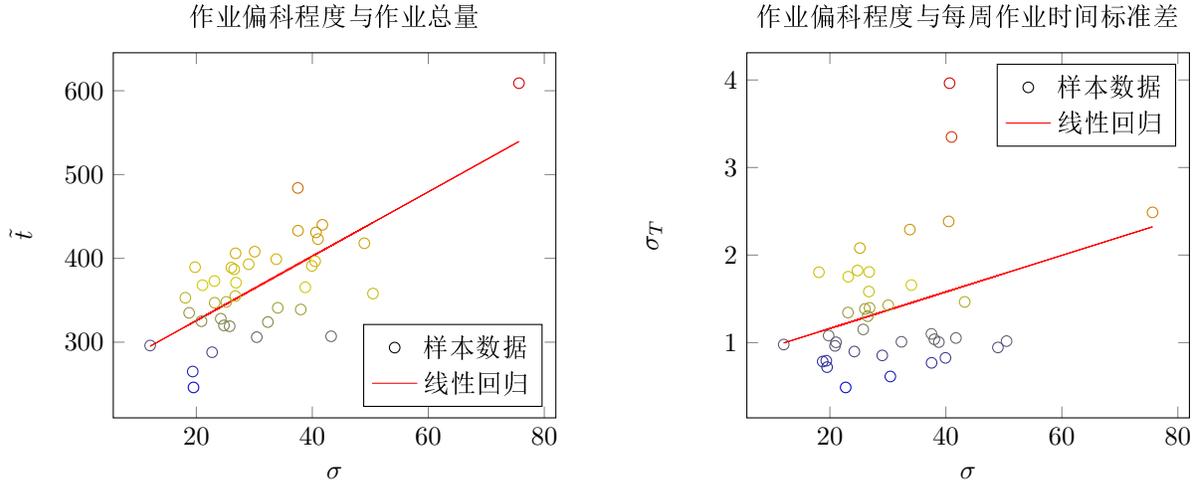
使用最小二乘法

$$\hat{\beta} = \arg \min |X\beta - Y|$$

得到

$$X^T X \hat{\beta} = X^T Y$$

图 4: 作业偏科程度与其他参量散点与线性回归图



如果矩阵 $X^T X$ 可逆, 则有唯一解

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (14)$$

解得

$$\begin{cases} a = -6.4189, \\ b = 0.3214, \\ c = -0.0036 \end{cases} \quad (15)$$

然后, 根据 Pearson 相关系数 (PCC) 的推论

$$\rho_{\Sigma \Delta D} = \sqrt{\frac{\sum_{j=1}^n (\widehat{\Delta d}_j - \overline{\Delta D})^2}{\sum_{j=1}^n (\Delta d_j - \overline{\Delta D})^2}} = 0.4117 > r_{0.10}(n-2) \quad (16)$$

所以在显著性水平 $\alpha = 0.10$ 下, 这个回归结果是显著的。

3.2 作业偏科程度与作业总时间、每周作业时间标准差

那么为什么这些作业偏科的人高考成绩会考得好呢? 我们首先考虑作业偏科程度 σ 与作业总时间 \bar{t} 的关系。采用同样的方法进行线性回归, 得到相关系数

$$\rho_{\Sigma \bar{t}} = 0.7033 > r_{0.10}(n-2) \quad (17)$$

所以在显著性水平 $\alpha = 0.10$ 下呈显著的正相关。

我们然后考虑作业偏科程度 σ 与每周作业时间标准差

$$\sigma_{T_j} = \sqrt{\frac{1}{9} \sum_{q=1}^9 (T_{qj} - \bar{T}_j)^2} \quad (18)$$

采用同样的方法进行线性回归, 得到相关系数

$$\rho_{\Sigma \sigma_T} = 0.3366 > r_{0.10}(n-2) \quad (19)$$

所以在显著性水平 $\alpha = 0.10$ 下线性回归显著。并呈正相关。

表 1: 每周每科作业时间正态估计参量与假设检验结果($\alpha = 0.05$)

i	学科	样本量	$\hat{\mu}_i$	$\hat{\sigma}_i$	χ_i^2	k	$\chi_\alpha^2(k-r-1)$	检验
1	CN	497	5.0413	2.8744	104.4943	7	9.4877	拒绝
2	MA	495	12.9101	5.9204	59.4031	9	12.59159	拒绝
3	EN	489	5.3957	2.4422	56.9568	5	5.9915	拒绝
4	PH	489	8.2873	3.3157	43.7089	8	11.0705	拒绝
5	CH	487	5.4825	2.6767	74.4935	6	7.814728	拒绝
6	BI	487	3.5452	2.4608	113.6225	5	5.9915	拒绝

4 每周每科作业时间服从正态分布的检验

最后用一个假设检验结尾。2017年10月, 我作出了一个假设: 我们收集到的数据中有每周每科的作业时间数据, 由于每次采集样本的时间间隔相同, 而且有长达十周的时间跨度, 我们可以认为这些数据的独立性等同, 虽然是同样的一批人做作业, 但是由于每次外界布置的作业是不同的, 可以被认为是不同的人的每周每科的作业时间, 也就意味着我们可以假设这些十周的每周每科作业时间数据相互独立。

在这一部分的分析中, 由于不牵涉到对同学差别的分析, 所以尽管有的同学有些每周每科作业时间数据缺失, 但是不影响对每周每科作业时间单变量分布的分析。

对于每周每科作业时间 $Q_i(i = 1, 2, 3, 4, 5, 6)$ 来说, 十周可以近似得到500个数据, 我们假设这些数据服从正态分布:

$$H_0: Q_i \sim N(\mu_i, \sigma_i^2) \quad (20)$$

首先进行参数估计, 运用矩估计法估计出参量:

$$\hat{\mu}_i = \bar{q}_i \quad (21)$$

$$\hat{\sigma}_i^2 = \frac{1}{n_0} \sum_{j=1}^{n_0} (q_{ij} - \bar{q}_i)^2 \quad (22)$$

然后为了利用非参数 χ^2 检验, 对于每个学科将取值离散化, 算出每一个取值组($q_{i,j-1} < Q_i \leq q_{ij}$)的频数 n_{ij} 和理论概率

$$\hat{p}_{ij} = P(q_{i,j-1} < Q_i \leq q_{ij}) = \Phi\left(\frac{q_{ij} - \hat{\mu}_i}{\hat{\sigma}_i}\right) - \Phi\left(\frac{q_{i,j-1} - \hat{\mu}_i}{\hat{\sigma}_i}\right) \quad (23)$$

令总频数

$$n_i = \sum_j n_{ij} \quad (24)$$

计算统计量 [, Appendix D]

$$\chi_i^2 = \sum_j \frac{(n_{ij} - n_i \hat{p}_{ij})^2}{n_i \hat{p}_{ij}} \quad (25)$$

定理 当假设 H_0 为真及 n_i 充分大时, 统计量 χ_i^2 总是近似服从自由度为 $k-r-1$ 的 χ^2 分布, 即

$$\chi_i^2 \sim \chi^2(k-r-1) \quad (26)$$

式中 k 为组数, 被估计参数个数 $r = 2$ 。

水平为 α 的拒绝域为

$$W = [\chi_\alpha^2(k-r-1), +\infty) \quad (27)$$

而这个定理的结论为近似结果, 应用时一般要求 $n \geq 50$, 且每个 $n_i p_{ij} \geq 5$, 否则相邻组要进行合并。

而表1显示在水平 $\alpha = 0.05$ 的情况下, 没有一个学科符合正态分布假设。这点推翻了2017年作出的符合正态分布的假设。图6显示了样本数据与假设结果的关系。

图 5: 每周每科作业量与正态分布假设散点图

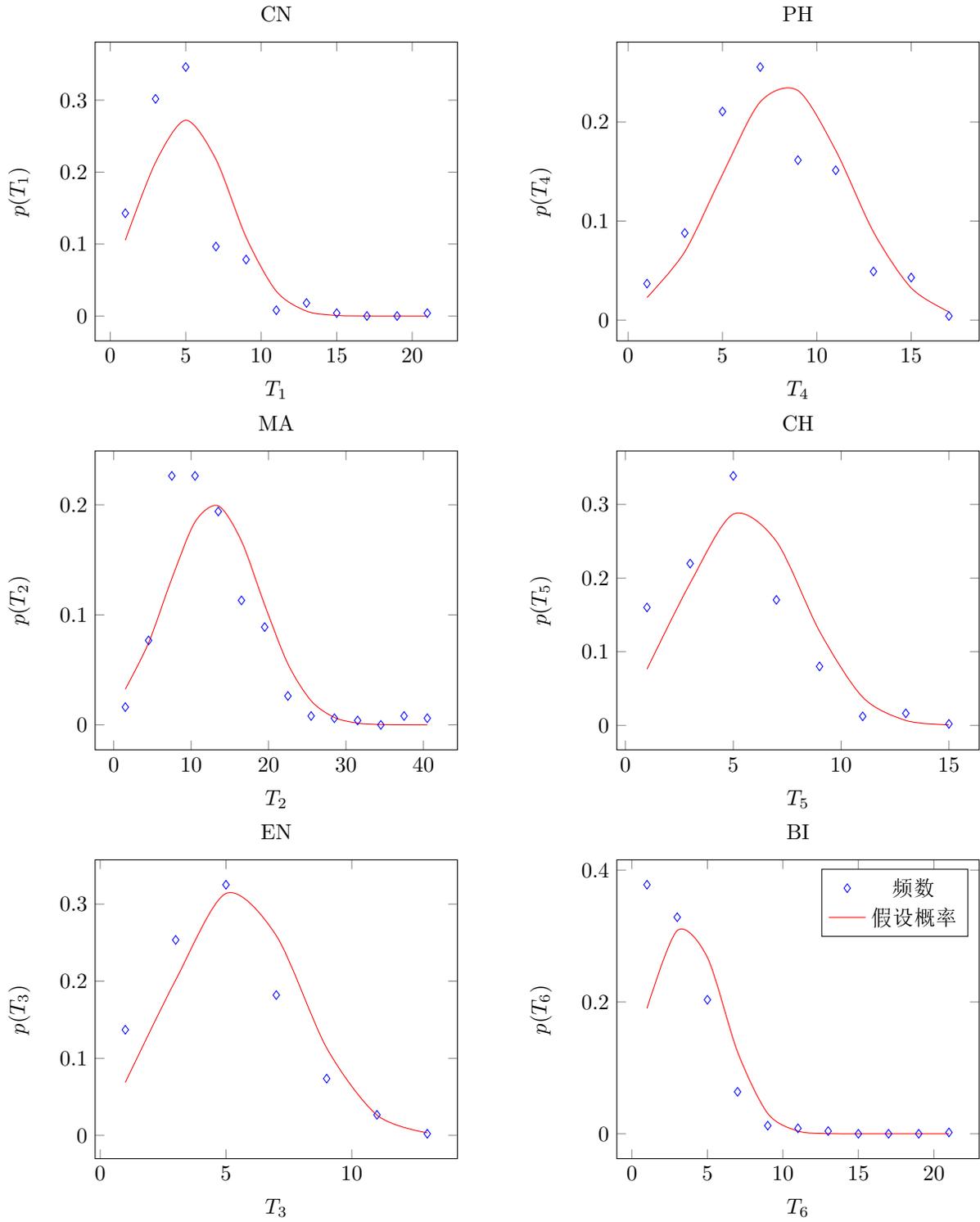


表 2: 回归情况表

回归组	节次	自变量	因变量	回归方式	ρ	$r_{\alpha}(n-2)^{[1]}$
I	2.1	作业总时间 $\tilde{T}^{[2]}$	高考成绩 S	线性	0.3066	
II	2.2	作业总时间 \tilde{T}	期末成绩提升效应 $\Delta D^{[3]}$	线性	-0.0236	
III	3.1	作业偏科程度 $\Sigma^{[4]}$	高考成绩 S	线性	0.4979	0.2707
IV	3.1	作业偏科程度 Σ	期末成绩提升效应 ΔD	二次 ^[5]	0.4117	
V	3.2	作业偏科程度 Σ	作业总时间 \tilde{t}	线性	0.7033	
VI	3.2	作业偏科程度 Σ	每周作业时间标准差 Σ_T	线性	0.3366	

¹ $\alpha = 0.10, n = 39$

² 9周半小时作业总时间。

³ 高一下学期至高二上学期期末考试成绩在全年级中偏差值的提升值。

⁴ 采用每周半小时作业时间的标准差表征。

⁵ 使用Pearson相关系数(PCC)的推论计算。

5 论证与结论

总时间与成绩 首先, 作业总量与高考成绩呈正相关(回归组I)而与期末成绩提升效应不明显相关(回归组II)。这意味着, 同学学习风格, 其作业量的多少, 与近期成绩的关系不大, 甚至可能会因为发挥失误而造成“付出了不一定有回报”的错觉。但缺少的实际上是一种毅力, 只要坚持到最后, 学习花的时间足够长, 就会在关键的时刻得到回报。

学科分配与成绩 其次, 我们看到, 近期而言(回归组IV), 在一段确定的时期内, 不是作业偏科程度越大越好, 过于偏科反而将导致负效应。当然, 如果作业偏科程度较小, 就意味着可能均匀用力, 但没有去专门的补自己的短板, 也会导致负效应。对于那些拥有自己学习方法的人, 他们会比较清楚自己的短板, 然后会“夹带私货”, 增加那个学科的学习时间, 从而不会“不进而退”, 反而“事半功倍”。

当然, 长期来看(回归组III), 这一段时间的作业偏科程度反应着个人的学习风格, 而与高考成绩呈现正相关意味着: 实际上这些作业偏科的人首先用更多的时间去学习(回归组V), 然后他们会形成自己的学习风格, 劳逸结合, 学习总时间更灵活一些(回归组VI), 这些都会给他们高考成绩加分。

正态分布 最后, 关于每周每科作业时间没有符合正态分布。原因大致是不能将九周的同一学生看成独立的个体, 并且一个班内作业布置作业情况相同, 独立性会有丧失。但是从图形来看, 已经比较接近了, 可能需要更多独立的个体调查, 或许就够得到正态分布的结果。如果能够得到正态分布的结果, 那么不仅会有研究价值, 而且对个人学科时间分配也会有很大的参考价值。

总之, 真正的学霸劳逸结合, 懂得自己的薄弱环节, 学足够长的时间, 并且推迟满足感, 才能在相对正态分布的群体中脱颖而出。

表 3: 评分参考表

代号	子项	\tilde{T}
BP	MA / <i>Others</i> ¹	2 / 1
3D(SP)		2
EP	CN,MA / <i>Others</i>	3 / 2
RT		0.5
XK		1.5
53	BI / PH	3 / 4.5
O(eg)	BI / <i>Others</i>	1 / 2
O(ex)	BI / MA / <i>Others</i>	1 / 3 / 2

¹ 科目有CN,MA,EN,PH,CH,BI。*Others*代表这个范围内的其他科目。

A 评分参考

参见表3。

B 计算细节

本文数据基本采用Microsoft Excel计算。

2.1节 F分位点 对于显著性水平 α ,第一自由度 m ,第二自由度 n ,F分布分位点 $F_\alpha(m, n)$

=F.INV(alpha,m,n)

2.2节 标准偏差 对于一组数据数据列D, 其标准偏差 σ

=STDEV.P(D:D)

4节 标准正态分布函数 需要求出(23)式的值, 对于 $O5=q_{ij-1}, P5=q_{ij}, M5=\hat{\mu}_i, N5=\hat{\sigma}_i$ 可以采用下面的公式计算

=NORM.DIST((P5-M5)/N5,0,1,TRUE)-NORM.DIST((O5-M5)/N5,0,1,TRUE)

4节 卡方分布分位点 合并组后的数据列E, 组数 k (在单元格F1)

=COUNT(E:E)

则 χ^2 分布在水平 α 下, 估计参量数 $r = 2$, 拒绝域起点 $\chi_\alpha^2(k - r - 1)$

=CHISQ.INV.RT(alpha,F1-3)

C 全年级期末成绩直方图

直方图参见图6。除去第一组数据0-34可能有一些缺考情况, 整体而言还是非常符合正态分布的。由于之前使用Khintchine大数定律已经说明了问题, 所以就不再假设检验。

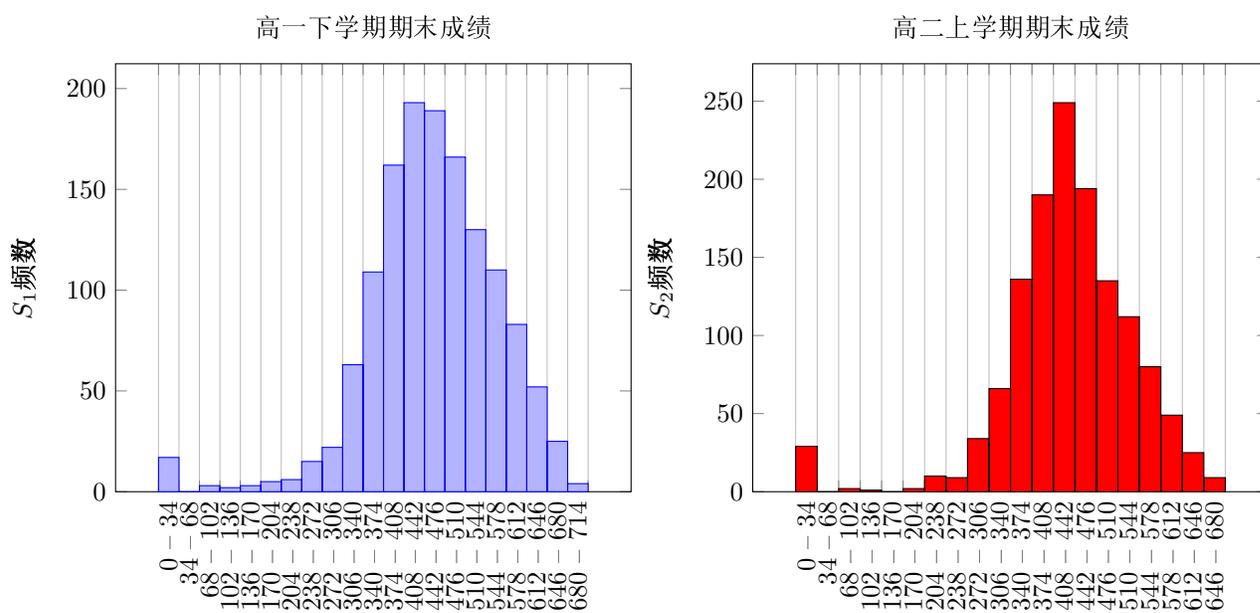
D 正态分布检验细节

参见表4。

表 4: 正态分布检验表

学科 i	总频数 n_i	均值 $\hat{\mu}_i$	方差 $\hat{\sigma}_i$	起点 q_{ij-1}	终点 q_{ij}	频数 n_{ij}	理论概率 \hat{p}_{ij}	$n_i \hat{p}_{ij}$	$n_{ij} - n_i \hat{p}_{ij}$	$\frac{(n_{ij} - n_i \hat{p}_{ij})^2}{n_i \hat{p}_{ij}}$
1	497	5.04	2.87	0	2	71	0.1053	52.3290	18.6710	6.6618
1	497	5.04	2.87	2	4	150	0.2136	106.1416	43.8584	18.1226
1	497	5.04	2.87	4	6	172	0.2721	135.2106	36.7894	10.0100
1	497	5.04	2.87	6	8	48	0.2177	108.1969	-60.1969	33.4914
1	497	5.04	2.87	8	10	39	0.1094	54.3762	-15.3762	4.3480
1	497	5.04	2.87	10	12	4	0.0345	17.1524	-13.1524	10.0852
1	497	5.04	2.87	12	14	9	0.0068	3.8472	9.1528	21.7753
1	497	5.04	2.87	14	16	2	0.0008			
1	497	5.04	2.87	16	18	0	0.0001			
1	497	5.04	2.87	18	20	0	0.0000			
1	497	5.04	2.87	20	22	2	0.0000			
104.4943										
2	495	12.91	5.92	0	3	8	0.0325	16.0726	-8.0726	4.0545
2	495	12.91	5.92	3	6	38	0.0745	36.8751	1.1249	0.0343
2	495	12.91	5.92	6	9	112	0.1329	65.7918	46.2082	32.4538
2	495	12.91	5.92	9	12	112	0.1844	91.2930	20.7070	4.6967
2	495	12.91	5.92	12	15	96	0.1990	98.5257	-2.5257	0.0647
2	495	12.91	5.92	15	18	56	0.1671	82.7014	-26.7014	8.6209
2	495	12.91	5.92	18	21	44	0.1091	53.9903	-9.9903	1.8486
2	495	12.91	5.92	21	24	13	0.0554	27.4116	-14.4116	7.5769
2	495	12.91	5.92	24	27	4	0.0219	10.8225	-6.8225	4.7996
2	495	12.91	5.92	27	30	3	0.0067	15.1084	0.8916	0.0526
2	495	12.91	5.92	30	33	2	0.0016			
2	495	12.91	5.92	33	36	0	0.0003			
2	495	12.91	5.92	36	39	4	0.0000			
2	495	12.91	5.92	39	42	3	0.0000			
59.4031										
3	489	5.40	2.44	0	2	67	0.0686	33.5588	33.4412	33.3241
3	489	5.40	2.44	2	4	124	0.2016	98.5982	25.4018	6.5442
3	489	5.40	2.44	4	6	159	0.3139	153.4862	5.5138	0.1981
3	489	5.40	2.44	6	8	89	0.2592	126.7262	-37.7262	11.2310
3	489	5.40	2.44	8	10	36	0.1134	69.8875	-19.8875	5.6593
3	489	5.40	2.44	10	12	13	0.0263			
3	489	5.40	2.44	12	14	1	0.0032			
56.9568										
4	489	8.29	3.32	0	2	18	0.0227	11.1219	6.8781	4.2536
4	489	8.29	3.32	2	4	43	0.0690	33.7572	9.2428	2.5307
4	489	8.29	3.32	4	6	103	0.1471	71.9546	31.0454	13.3948
4	489	8.29	3.32	6	8	125	0.2203	107.7411	17.2589	2.7647
4	489	8.29	3.32	8	10	79	0.2318	113.3450	-34.3450	10.4070
4	489	8.29	3.32	10	12	74	0.1713	83.7782	-9.7782	1.1413
4	489	8.29	3.32	12	14	24	0.0890	43.5026	-19.5026	8.7432
4	489	8.29	3.32	14	16	21	0.0324	19.9275	3.0725	0.4737
4	489	8.29	3.32	16	18	2	0.0083			
43.7089										
5	487	5.48	2.68	0	2	78	0.0763	37.1820	40.8180	44.8095
5	487	5.48	2.68	2	4	107	0.1932	94.0958	12.9042	1.7697
5	487	5.48	2.68	4	6	165	0.2868	139.6793	25.3207	4.5901
5	487	5.48	2.68	6	8	83	0.2499	121.6910	-38.6910	12.3016
5	487	5.48	2.68	8	10	39	0.1277	62.2107	-23.2107	8.6599
5	487	5.48	2.68	10	12	6	0.0383	22.2508	-7.2508	2.3628
5	487	5.48	2.68	12	14	8	0.0067			
5	487	5.48	2.68	14	16	1	0.0007			
74.4935										
6	487	3.55	2.46	0	2	184	0.1902	92.6221	91.3779	90.1504
6	487	3.55	2.46	2	4	160	0.3083	150.1374	9.8626	0.6479
6	487	3.55	2.46	4	6	99	0.2674	130.2431	-31.2431	7.4947
6	487	3.55	2.46	6	8	31	0.1241	60.4461	-29.4461	14.3446
6	487	3.55	2.46	8	10	6	0.0308	17.1046	-4.1046	0.9850
6	487	3.55	2.46	10	12	4	0.0041			
6	487	3.55	2.46	12	14	2	0.0003			
6	487	3.55	2.46	14	16	0	0.0000			
6	487	3.55	2.46	16	18	0	0.0000			
6	487	3.55	2.46	18	20	0	0.0000			
6	487	3.55	2.46	20	22	1	0.0000			
113.6225										

图 6: 全年级期末考试成绩频数直方图



E 致谢

感谢高中班主任陈崇洲老师提供高考成绩、期末考试成绩以及相关支持。感谢班委杜浩宇、刘湘宁等12名班委执行统计活动(2015 - 2017)。特别感谢谷文治同学的支持。

References

- [1] 暴风计划技术性报告, LogCreative, 2017.3
- [2] 风华班作业量分析, LogCreative, 2017.10
- [3] 偏差值, 百度百科, <https://baike.baidu.com/item/偏差值/6096041?fr=aladdin>
- [4] 概率论与数理统计, 武爱文等, 2版——上海:上海交通大学出版社, 2014
- [5] Introduction to Linear Algebra, Gilbert Strang, Fifth Edition, Wellesley-Cambridge Press, 2016
- [6] Pearson correlation coefficient, Wikipedia, https://en.wikipedia.org/wiki/Pearson_correlation_coefficient