基于日步数与睡眠时长的个人行为数据统计分析报告

姓名：何子宁　学号：524030910131

2025 年 12 月

摘要

本文基于运动手环记录的 90 天日度数据（步数与睡眠时长）进行统计分析。首先给出两变量的描述性统计与可视化（时间序列、分布图），随后对步数采用对数正态近似、对睡眠时长采用正态近似，完成参数估计与 95% 置信区间，并以“周末与工作日步数是否不同”“步数与睡眠是否相关”为例给出假设检验过程，最后报告线性回归结果。样本中日步数均值为 8500.93 步、波动较大，睡眠时长均值为 7.26 小时；相关性检验得到 $r = 0.518$（$p \approx 1.71 \times 10^{-7}$），回归中睡眠时长的斜率估计为正（约 1775 步/小时）。结论主要用于说明方法应用，结果仍可能受极端值与其他未纳入因素影响。

1　引言

日步数与睡眠时长是个人健康管理中常用的两个日度指标：步数反映活动水平，睡眠时长反映恢复与作息。以 90 天的日度记录为样本，可以直观看到波动与分布特点，也能按照课堂方法把概率建模、估计、检验与回归的步骤完整做一遍。

本报告聚焦以下问题：

1. 日步数与睡眠时长的总体水平与离散程度如何？是否存在异常值或偏态？
2. 两变量在 90 天内的日度波动与趋势如何？是否存在明显的阶段性变化？
3. 日步数与睡眠时长是否存在统计相关？睡眠时长对步数的线性关联强度有多大？

2　数据集基本介绍

2.1　数据来源

数据来自运动手环记录的每日汇总值，按日口径整理得到 90 天的日步数与睡眠时长序列（起始日期为 2025-09-16，截止日期为 2025-12-14）。

2.2 变量与类型

- 日期：t（按日时间索引）
- 日步数：Y_t（离散、非负计数型）
- 睡眠时长：S_t（连续型，单位：小时）

为辅助描述时间结构与趋势，可由日期推导衍生变量：

- 是否周末：$W_t \in \{0,1\}$（周末为 1，工作日为 0）
- 7 日平均步数/睡眠时间：$\overline{Y}_t = \frac{1}{7}\sum_{i=0}^{6} Y_{t-i}$，$\overline{S}_t = \frac{1}{7}\sum_{i=0}^{6} S_{t-i}$

2.3 数据规模

原始核心变量维数为 2（Y_t, S_t），样本量为 $n = 90$。用于可视化与检验的 $W_t, \overline{Y}_t, \overline{S}_t$ 为由日期或滑动窗口计算得到的衍生变量。

2.4 数据表结构

表1展示数据表结构与样例记录；完整 90 天数据见附录表7。

日期	日步数 Y_t	睡眠时长 S_t (h)	周末 W_t	备注
2025-09-16	812	2.3	0	
2025-09-17	1247	6.4	0	
2025-09-18	1793	4.1	0	

表 1: 按日数据记录格式与样例

2.5 数据预处理

- 测量范围统一：步数与睡眠均为按日汇总值，日期连续覆盖 90 天。
- 缺失值：本样本两变量均无缺失。
- 异常值：步数存在极低值（最低 812 步），睡眠也存在极端值（最低 2.3 小时、最高 13.2 小时）。分析过程中保留原始记录，并在结论中讨论其可能含义与对统计分析的影响。

2.6 衍生变量

为刻画短期噪声与趋势，计算步数与睡眠的 7 日移动平均，并在图中与原始序列叠加展示。

3 可视化分析

3.1 总体描述

表2给出两变量的描述性统计。均值与标准差反映总体水平与离散程度；考虑到存在极端值，同时报告中位数等稳健统计量。样本期内：

- 日步数均值为 8500.93 步、标准差为 5421.00 步，中位数为 6653 步；最小值为 812 步（2025-09-16），最大值为 21836 步（2025-11-15）。
- 睡眠时长均值为 7.26 小时、标准差为 1.58 小时，中位数为 7.2 小时；最小值为 2.3 小时（2025-09-16），最大值为 13.2 小时（2025-11-15）。

变量	均值	标准差	最小值	中位数	最大值
日步数 Y_t	8500.93	5421.00	812	6653	21836
睡眠时长 S_t	7.26	1.58	2.3	7.2	13.2

表 2: 描述性统计汇总

3.2 时间序列图与周期性

图1展示日步数与睡眠时长的时间序列及各自的 7 日移动平均。从图中可见步数的波动幅度明显大于睡眠时长；叠加移动平均后，部分时段两者变化方向较一致。

4 数据建模

4.1 (a) 概率模型、参数估计与区间估计

考虑到步数为非负且分布右偏明显，本文用对数正态分布对步数作近似；睡眠时长为连续变量，采用正态分布近似，便于进行参数估计与区间估计。

- 日步数 Y_t 为非负且右偏明显，采用对数正态模型近似：$Y_t \sim \mathrm{LogNormal}(\mu_Y, \sigma_Y^2)$，等价于 $Z_t = \ln Y_t \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$。
- 睡眠时长 S_t 为连续变量，采用正态模型：$S_t \sim \mathcal{N}(\mu_S, \sigma_S^2)$。

在独立同分布假设下，对数正态模型的极大似然估计为

$$\hat{\mu}_Y = \frac{1}{n}\sum_{t=1}^{n} \ln Y_t, \qquad \hat{\sigma}_Y^2 = \frac{1}{n}\sum_{t=1}^{n} (\ln Y_t - \hat{\mu}_Y)^2.$$

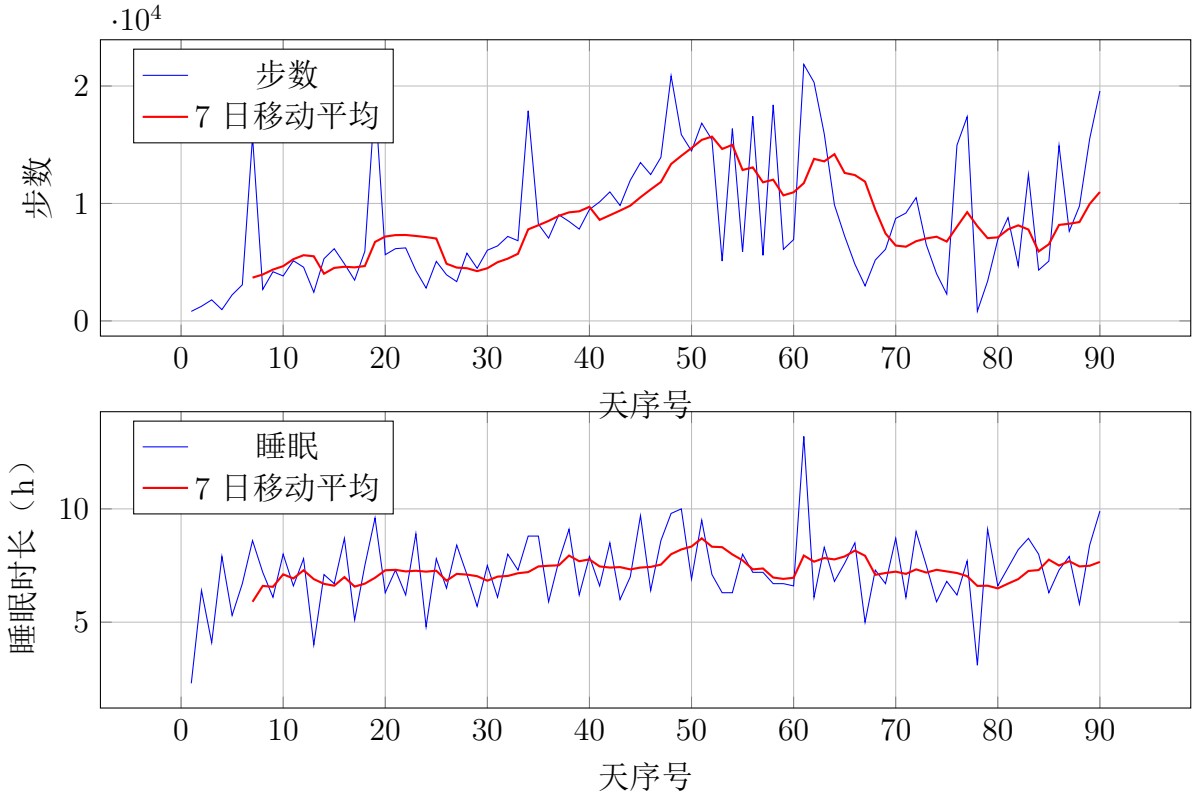

图 1: 日步数与睡眠时长的时间序列及 7 日移动平均

由样本计算得到 $\hat{\mu}_Y = 8.817$、$\hat{\sigma}_Y = 0.733$；对应的模型隐含步数均值为 $\exp(\hat{\mu}_Y + \frac{1}{2}\hat{\sigma}_Y^2) = 8833.32$（步）。对 Z_t 做区间估计得到

$$\mu_Y \in [8.666,\ 8.969], \qquad \sigma_Y^2 \in [0.410,\ 0.737].$$

在正态模型 $S_t \sim \mathcal{N}(\mu_S, \sigma_S^2)$ 下，μ_S 可用样本均值 $\bar{S}$ 估计，σ_S^2 可由样本方差 s_S^2 刻画。样本计算得到 $\bar{S} = 7.258$（h），$s_S = 1.581$（h）。进一步由 t 分布与 χ^2 分布可构造 95% 置信区间：

$$\mu_S \in [6.927,\ 7.589], \qquad \sigma_S^2 \in [1.903,\ 3.436].$$

4.2 (b) 二维散点图与线性回归

取二维随机变量 $(X_t, Y_t) = (S_t, Y_t)$。图3展示散点图与最小二乘拟合直线。

建立线性回归模型

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \quad \mathbb{E}(\varepsilon_t \mid X_t) = 0,$$

并进一步加入周末指示 W_t 作对比：

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 W_t + \varepsilon_t.$$

表3给出估计结果。模型 (1) 的 $R^2 = 0.268$，表明线性模型可解释约 26.8% 的步数波动；睡眠时长系数为正且显著。加入周末控制后，睡眠时长系数仍为正且显著，而周末项不显著。

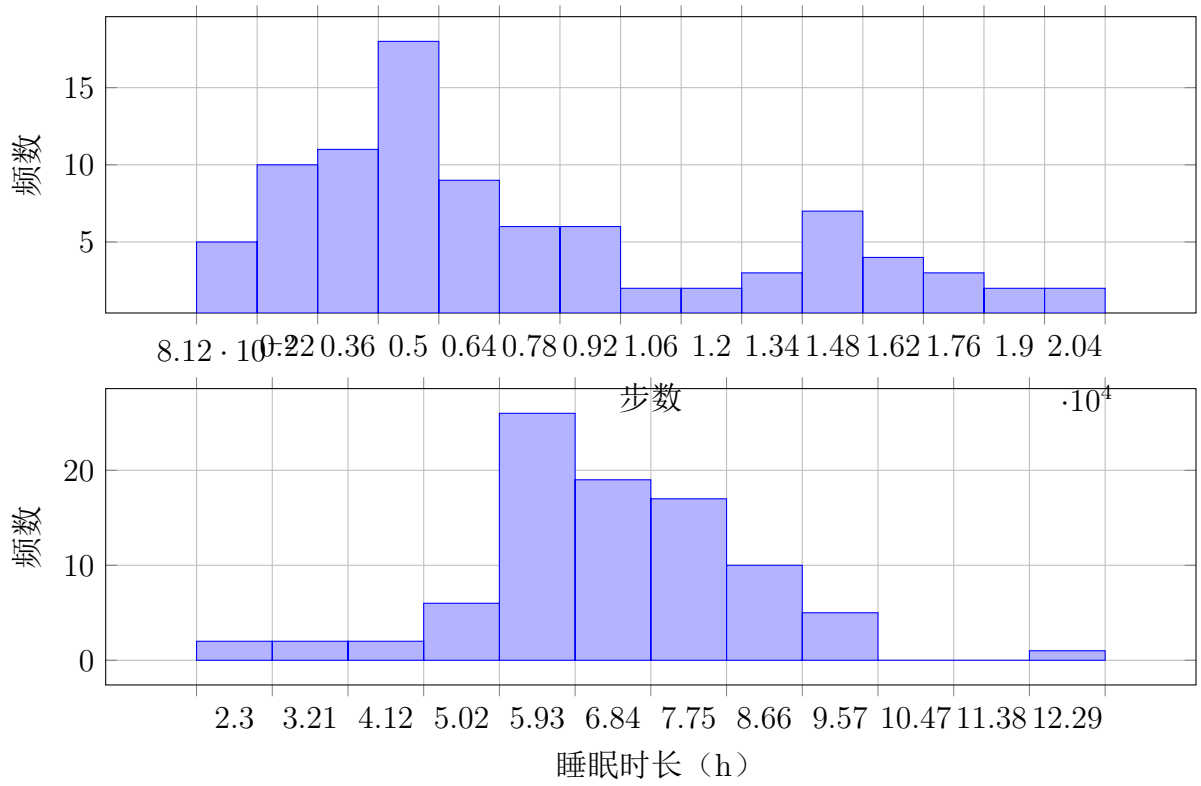

图 2: 日步数与睡眠时长的分布直方图

	模型 (1)：$Y \sim S$			模型 (2)：$Y \sim S + W$		
变量	系数估计	标准误差	p 值	系数估计	标准误差	p 值
截距 β_0	-4385.06	2321.18	0.062			
睡眠时长 β_1	1775.47	312.57	1.71×10^{-7}	1699.47	315.05	5.85×10^{-7}
周末指示 β_2				1581.44	1093.18	0.152

表 3: 线性回归结果（因变量：日步数）

5 假设检验

5.1 检验 1：周末与工作日步数差异

为检验步数是否存在显著的“周末效应”，将样本按 W_t 分为工作日组与周末组，比较两组步数均值是否相同。

$$H_0 : \mu_{\text{weekend}} = \mu_{\text{weekday}}, \quad H_1 : \mu_{\text{weekend}} \neq \mu_{\text{weekday}}$$

Welch 两样本 t 检验的统计量可写为

$$T = \frac{\bar{Y}_{\text{weekend}} - \bar{Y}_{\text{weekday}}}{\sqrt{s^2_{\text{weekend}}/n_{\text{weekend}} + s^2_{\text{weekday}}/n_{\text{weekday}}}},$$

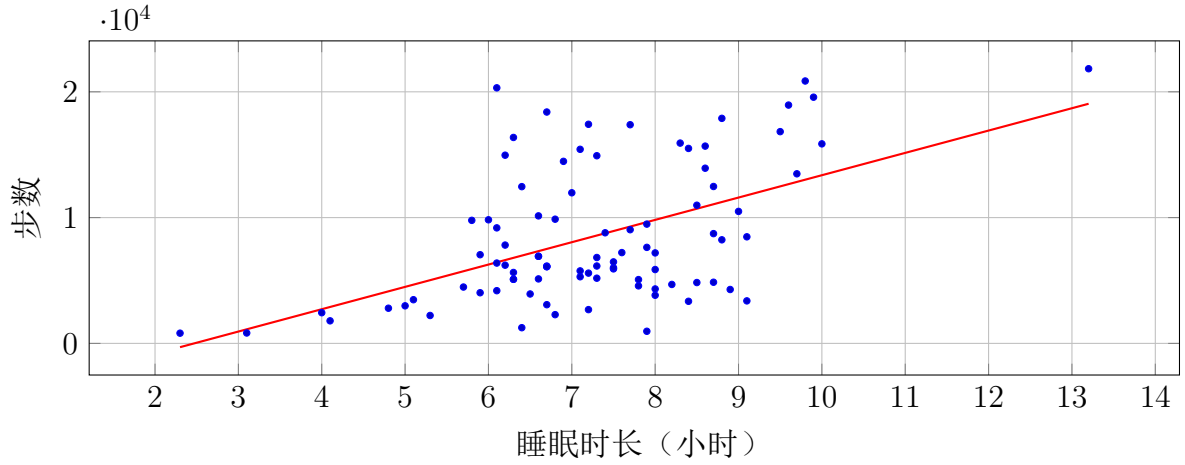

图 3: 日步数与睡眠时长散点图

其在 H_0 下近似服从自由度为 Welch–Satterthwaite 公式给出的 t 分布；在显著性水平 $\alpha = 0.05$ 下，拒绝域为 $|T| > t_{0.975,\mathrm{df}}$。同时报告 Mann–Whitney 秩和检验（其统计量经标准化后在大样本下近似服从 $\mathcal{N}(0,1)$，拒绝域为 $|Z| > 1.96$）。

5.2 结果呈现

表4与表5给出两组的描述统计与检验结果。周末步数均值高于工作日（10324.77 vs 7760.00），但在常用显著性水平下差异不显著。

组别	样本量	均值	标准差	中位数
工作日	64	7760.00	4598.54	6432.5
周末	26	10324.77	6816.48	8155

表 4: 工作日/周末的步数统计

检验方法	统计量	p 值
两样本 t 检验（Welch）	$t = 1.763$（df=34.63）	0.087
秩和检验（Mann–Whitney）	$z = -1.166$	0.244

表 5: 组间差异检验结果

5.3 检验 2：步数与睡眠时长的相关性

将“是否存在线性相关”表述为对总体相关系数的检验：

$$H_0 : \rho = 0, \quad H_1 : \rho \neq 0.$$

在 H_0 下，检验统计量

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

服从自由度为 $n-2$ 的 t 分布。给定显著性水平 $\alpha = 0.05$，拒绝域为 $|T| > t_{0.975,n-2}$；也可使用 p 值规则。由样本计算得到 $r = 0.518$（95% CI：$[0.348, 0.655]$），对应 $t = 5.680$（df=88），$p \approx 1.71 \times 10^{-7}$，据此拒绝 H_0。若检验不显著，则说明在该样本下没有足够证据支持线性相关。散点图见图3。

5.3.1 结果呈现

r	95% CI	t（df）	p 值	结论（$\alpha = 0.05$）
0.518	$[0.348, 0.655]$	5.680（88）	1.71×10^{-7}	拒绝 H_0

表 6: 相关性检验结果（$H_0 : \rho = 0$）

6 结论

- 本文依次完成数据集介绍、可视化、概率建模与参数估计、区间估计、假设检验与线性回归，并给出对应的计算结果与解释。
- 在 2025-09-16 至 2025-12-14 的 90 天样本中，日步数均值为 8500.93 步、标准差为 5421.00 步，中位数为 6653 步，存在极低步数日（812 步）；睡眠时长均值为 7.26 小时、标准差为 1.58 小时，中位数为 7.2 小时，亦存在极端睡眠日（2.3 小时与 13.2 小时）。
- 两变量关系方面，日步数与睡眠时长呈中等强度正相关（Pearson $r = 0.518$，95% CI：$[0.348, 0.655]$；Spearman $\rho = 0.409$）。线性回归结果显示：睡眠时长每增加 1 小时，日步数平均增加约 1775 步，且在统计意义上显著。
- 补充分析表明周末步数均值高于工作日，但差异未达到常用显著性水平；在回归中加入周末控制后，睡眠对步数的正向关联仍显著，而周末项不显著。
- 本文只基于两个核心变量做分析，且数据为单人记录；步数与睡眠都容易受当天安排、设备记录等影响，因此结果更适合作为相关性与方法示例，而非因果结论。

7 参考文献

参考文献

[1] 卫淑芝, 熊德文, 皮玲. 概率论与数理统计 [M]. 高等教育出版社, 2020

A 附录：完整数据

表 7: 完整数据（90 天）

天序号	步数	睡眠时长 (h)	周末
1	812	2.3	0
2	1,247	6.4	0
3	1,793	4.1	0
4	963	7.9	0
5	2,218	5.3	1
6	3,076	6.7	1
7	15,684	8.6	0
8	2,684	7.2	0
9	4,189	6.1	0
10	3,827	8	0
11	5,126	6.6	0
12	4,573	7.8	1
13	2,439	4	1
14	5,297	7.1	0
15	6,148	6.7	0
16	4,861	8.7	0
17	3,472	5.1	0
18	5,924	7.5	0
19	18,946	9.6	1
20	5,638	6.3	1
21	6,152	7.3	0
22	6,217	6.2	0
23	4,284	8.9	0
24	2,796	4.8	0
25	5,071	7.8	0
26	3,928	6.5	1
27	3,346	8.4	1
28	5,763	7.1	0
29	4,479	5.7	0
30	6,018	7.5	0

续下页

表 7: 完整数据（90 天，续）

天序号	步数	睡眠时长 (h)	周末
31	6,386	6.1	0
32	7,193	8	0
33	6,827	7.3	1
34	17,891	8.8	1
35	8,234	8.8	0
36	7,049	5.9	0
37	9,037	7.7	0
38	8,472	9.1	0
39	7,816	6.2	0
40	9,483	7.9	1
41	10,137	6.6	1
42	10,984	8.5	0
43	9,826	6	0
44	11,973	7	0
45	13,482	9.7	0
46	12,461	6.4	0
47	13,927	8.6	1
48	20,861	9.8	1
49	15,863	10	0
50	14,472	6.9	0
51	16,834	9.5	0
52	15,429	7.1	0
53	5,094	6.3	0
54	16,372	6.3	1
55	5,871	8	1
56	17,418	7.2	0
57	5,587	7.2	0
58	18,394	6.7	0
59	6,097	6.7	0
60	6,924	6.6	0
61	21,836	13.2	1
62	20,317	6.1	1
63	15,924	8.3	0
64	9,874	6.8	0
65	7,218	7.6	0
66	4,837	8.5	0
67	2,986	5	0
68	5,183	7.3	1
69	6,097	6.7	1
70	8,726	8.7	0

续下页

表 7: 完整数据（90 天，续）

天序号	步数	睡眠时长 (h)	周末
71	9,184	6.1	0
72	10,493	9	0
73	6,479	7.5	0
74	4,026	5.9	0
75	2,284	6.8	1
76	14,962	6.2	1
77	17,384	7.7	0
78	827	3.1	0
79	3,379	9.1	0
80	6,924	6.6	0
81	8,793	7.4	0
82	4,686	8.2	1
83	12,473	8.7	1
84	4,327	8	0
85	5,094	6.3	0
86	14,918	7.3	0
87	7,634	7.9	0
88	9,781	5.8	0
89	15,500	8.4	1
90	19,573	9.9	1