東南大學

数据统计分析

课程报告

2023-2024 赛季 NBA 球员关键数据

基于 SAS 的统计分析

项目成员: 陈奕诚

张韫译萱

周一鸣

指导老师: 薛磊

2025 年 11 月 25 日

目录

1 选题背景与数据预处理

1.1 研究背景

美国职业篮球联赛 NBA 是目前世界上竞技水平最高、商业化程度和全球影响力都非常突出的职业篮球联赛之一。常规赛阶段包括 30 支球队、每队 82 场比赛，会产生数量庞大的技术统计数据，涵盖球员个人表现、球队整体效率以及多种高级指标，数据量大、结构清晰、记录规范，非常适合作为统计分析与建模的数据集。

从统计学教学的角度来看，NBA 数据有以下几个方面的优势。

1. 数据量充足且具有一定代表性。单赛季累计记录数以万计，覆盖数百名球员及其多维度表现，在样本容量上能够满足描述统计与假设检验的需求。
2. 规则明确、指标统一。比赛规则和技术统计指标高度标准化，不同球队、不同球员的数据可以直接横向比较，有利于聚焦统计方法本身，而不被数据口径差异干扰。
3. 直观性强、便于解释。得分、助攻、篮板、盖帽、抢断等指标与实际比赛情景高度对应，方便将统计结论与战术位置、角色分工联系起来，提高分析结果的可解释性与趣味性。

在本次课程设计中，小组选择 2023–2024 赛季常规赛球员数据作为研究对象，重点围绕球员场上位置与若干关键技术指标之间的关系展开分析。具体而言，关注后卫与锋线或内线球员在得分、助攻、篮板、盖帽和抢断等维度上的差异，并利用非参数统计方法检验这些差异是否显著。

1.2 数据来源与变量说明

本次分析所用原始数据为 `2023-2024 NBA Player Stats –Regular.csv`。每一行对应一名球员在该赛季常规赛的场均统计数据，包括出场次数、场均上场时间、多类投篮命中率以及若干高级效率指标等。

在众多变量中，本报告重点关注以下几类信息。

1. 基本信息类。
   (a) 球员姓名 Player。
   (b) 场上位置 Pos。
   (c) 年龄 Age。
   (d) 所在球队 Tm。

(e) 出场场数 G。

(f) 场均上场时间 MP。

2. 进攻表现类。

(a) 场均得分 PTS。

(b) 场均助攻 AST。

(c) 投篮命中率 FG%。

(d) 三分命中率 3P%。

(e) 罚球命中率 FT%。

(f) 有效投篮命中率 eFG%。

3. 篮板与防守类。

(a) 场均总篮板 TRB。

(b) 场均抢断 STL。

(c) 场均盖帽 BLK。

4. 失误及其他指标。

(a) 场均失误 TOV。

(b) 其他高级效率指标，本次分析未列为重点。

考虑到课程设计的重点在于统计方法的运用与解释，在变量数量较多的情况下，本报告选取了 5 个最具代表性的关键技术指标，即 PTS、AST、TRB、BLK、STL 作为后续分析的核心对象。

1.3 样本筛选与预处理原则

原始数据覆盖 700 余名球员，其中包括部分出场时间较少、数据波动较大的边缘球员。如果将这些样本直接纳入统计分析，可能导致结论不稳定，同时会削弱位置与技术指标之间关系的可解释性。

为保证样本的稳定性与可比性，本项目采用了如下筛选原则。

1. 出场数筛选。仅保留出场次数 $G \geq 20$ 的球员，排除赛季中短合同、长期伤停或仅打少量比赛的特殊样本。

2. 上场时间筛选。要求场均上场时间 $MP \geq 15$ 分钟，保证样本球员在球队中真正承担轮换甚至主力角色，其赛季平均数据具有统计意义。

3. 变量选择与重命名。在 30 多个技术指标中筛选出用于本次分析的核心指标，对个别变量进行简要重命名或类型转换，以便在 SAS 中处理。

经过上述筛选操作后，样本量缩减为原始数据的大约 1/4，但保留了各队主要轮换球员的信息，兼顾样本数量与数据质量，为后续分析打下基础。

1.4 位置分组变量的构造

本次分析的重点是比较不同位置球员在关键技术指标上的表现差异。考虑到 NBA 官方位置划分较为细致，且不少球员兼任多个位置，如果逐个位置进行比较，会出现样本量不足或者解释困难的问题。

为此，本项目按照篮球战术分工的常见做法，将位置简化为两大类。

1. 后卫 Guard。只要 Pos 字段中包含字母 G，如 PG、SG、G-F 等，即视为后卫类球员。这类球员一般负责持球推进、组织进攻和外线投射。

2. 前场 Frontcourt。其他所有位置如 F、C、F-C 等统一归为前场球员，一般承担篮板保护、内线得分和护筐任务。

在 SAS 中通过构造新的分组变量 `Group` 表示以上划分,取值为 `Guard` 或 `Frontcourt`，以便后续进行描述统计和非参数检验。

2 数据分析

2.1 分析思路与方法概述

本次课程设计的核心问题可以概括为下面这一点。

在 2023–2024 赛季主要轮换球员中，后卫和前场球员在 5 项关键技术指标上是否存在显著差异。

围绕上述问题，本报告采用“描述统计 + 可视化 + 非参数检验”的基本框架。

1. 对 5 个关键指标 PTS、AST、TRB、BLK、STL 计算均值、方差、标准差、变异系数、偏度、峰度、中位数、上下四分位数和四分位极差等描述统计量，并通过直方图和箱线图进行可视化展示。

2. 以位置分组变量 Group 为自变量，使用 Wilcoxon 秩和检验以及 Kruskal–Wallis 检验对 5 个指标进行双样本比较，检验位置对关键技术指标的影响是否具有统计显著性。

3. 结合统计结果与篮球战术常识，对差异的实际意义进行解读，得出相应结论。

2.2 SAS 实现流程与关键输出

在具体实现层面，小组使用 SAS 完成了从数据导入到结果输出的完整流程。核心代码结构和与之对应的输出可以概括为以下几个步骤。

1. 数据导入与变量预处理。使用 `PROC IMPORT` 将原始 CSV 文件读入数据集，设置 `options validvarname=any` 以允许变量名中出现百分号等字符，并在数据步中构造中文姓名变量和分组变量 `Group`。

2. PTS 整体描述统计。使用 `PROC MEANS` 对 PTS 进行整体描述统计，输出样本量 N、均值 Mean、方差 Var、标准差 Std、变异系数 CV、偏度 Skewness、峰度 Kurtosis、中位数 Median、一四分位数 $Q1$、三四分位数 $Q3$、四分位极差 IQR 等指标。该输出在 PPT 左侧展示，本报告中以图 1 示意。

3. PTS 分布形态分析。使用 `PROC UNIVARIATE` 生成带正态拟合和核密度估计的直方图，同时通过 ODS 输出 Moments、Quantiles 与 BasicMeasures 三类统计表。直方图与拟合曲线的图像示意见图 2。

4. 箱线图与分组描述统计。使用 `PROC SGPLOT` 绘制 PTS 的总体箱线图和按 Group 分组的箱线图，并对 Guard 与 Frontcourt 两组分别计算描述统计量。箱线图示意见图 3。对 AST、TRB、BLK、STL 同样绘制按位置分组的箱线图，分别见图 4 和图 5。

PTS 整体描述统计的 SAS 输出示意如图 1 所示。

统计量	数值
样本量	708
算术平均数	12.65
中位数	11.50
众数	10.00
方差	30.81
标准差	5.55
变异系数	43.87%
偏度	0.89
峰度	0.73
下四分位数（Q1）	8.70
上四分位数（Q3）	15.40
四分位极差（IQR）	6.70

图 1: SAS 输出的 PTS 整体描述统计示意

2.3 描述统计与总体分布概览

在完成数据筛选后，首先对全部轮换球员样本的场均得分 PTS 进行整体描述统计，并绘制带正态拟合与核密度估计的直方图。直方图示意如图 2 所示。

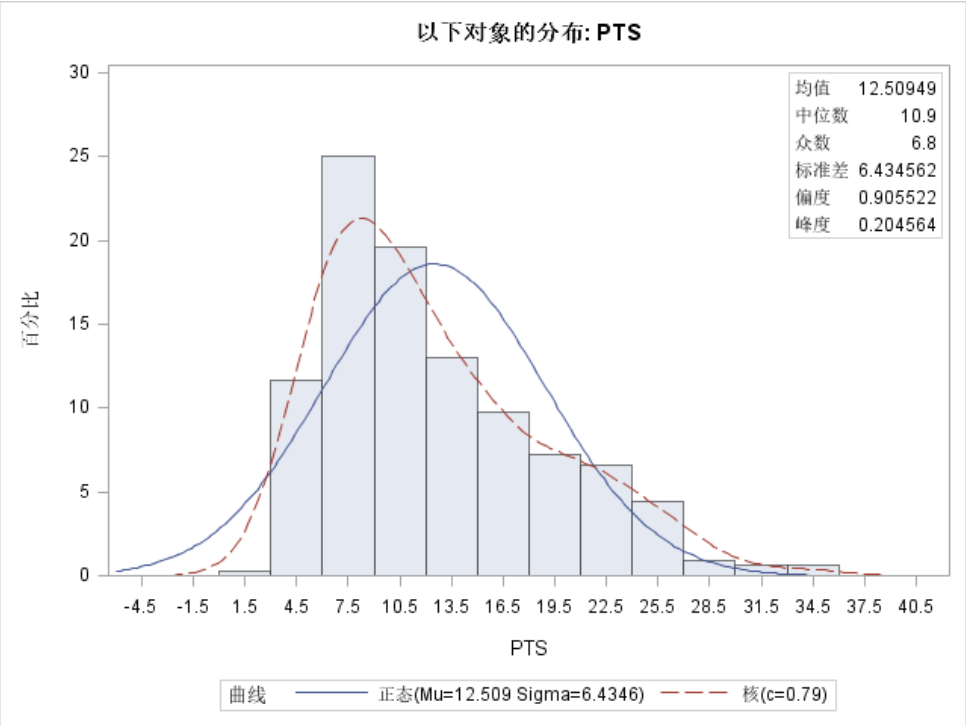

图 2: 主要轮换球员场均得分直方图与拟合曲线示意

从直方图和拟合曲线可以看出以下几点。

1. PTS 的分布大致呈右偏形态。多数球员场均得分集中在 10 分左右，极少数高产得分手形成较长的右尾。

2. 偏度和峰度的数值进一步量化了非对称性和尖峭程度，与直方图的观察结果基本一致。

3. 中位数与四分位数表明，约有 50% 的轮换球员场均得分落在一个相对集中的区间内，对异常高分球员不敏感，更适合作为“典型表现”的刻画。

进一步地，对 PTS、AST、TRB、BLK、STL 分别绘制总体及按位置分组的箱线图，用于直观比较后卫与前场在不同指标上的分布差异。场均得分 PTS 的总体与分组箱线图示意见图 3，AST 与 TRB 的位置箱线图示意见图 4，BLK 与 STL 的位置箱线图示意见图 5。

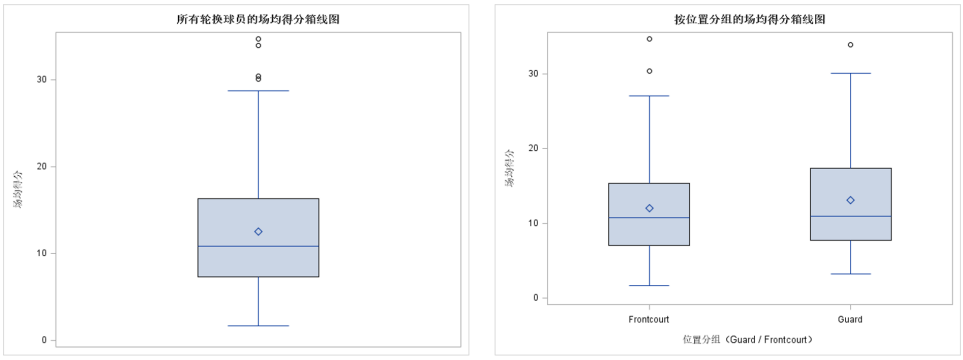

图 3: 场均得分 PTS 的总体与按位置分组箱线图示意

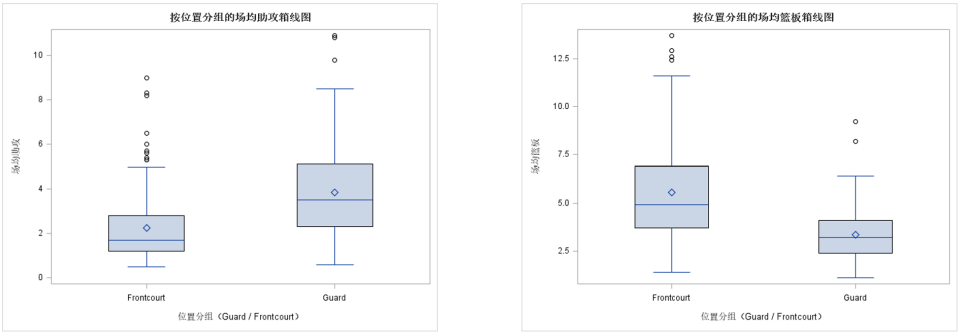

图 4: 按位置分组的助攻 AST 与篮板 TRB 箱线图示意

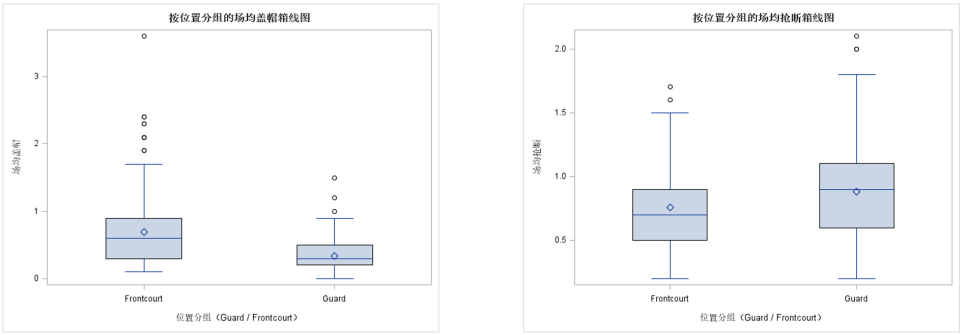

图 5: 按位置分组的盖帽 BLK 与抢断 STL 箱线图示意

2.4 Wilcoxon 秩和检验与 Kruskal–Wallis 检验方法说明

由于 5 个关键指标的分布并不完全满足正态性假设，且不同位置组的方差可能存在差异，如果直接采用 t 检验可能会导致结论不稳健。本项目因此采用 Wilcoxon 秩和检验和 Kruskal–Wallis 检验两种非参数方法对两组位置球员进行比较。

2.4.1 Wilcoxon 秩和检验

以某一技术指标 X 为例，设后卫组样本来自总体 X_G，前场组样本来自总体 X_F，检验问题可以写为

$$H_0 : X_G \text{ 与} X_F \text{ 的分布无显著差异} \quad \text{对比} \quad H_1 : X_G \text{ 与} X_F \text{ 的分布存在差异.} \tag{1}$$

Wilcoxon 秩和检验的基本思想是将两组样本合并后按数值大小排序，计算其中一组样本的秩和并与原假设成立时的期望秩和进行比较。当样本量足够大时，秩和统计量在适当标准化后近似服从正态分布，从而可以得到标准化的 Z 统计量及其对应的双侧 p 值。在 SAS 的 `PROC NPAR1WAY` 输出中，包含秩和、期望秩和、标准误以及 Z 值和双侧 p 值等信息。

2.4.2 Kruskal–Wallis 检验

Kruskal–Wallis 检验是基于秩的一种多组比较方法，用于检验若干总体分布是否相同。在本项目中仅有两组（Guard 与 Frontcourt），此时 Kruskal–Wallis 检验在本质上与 Wilcoxon 秩和检验等价，但其统计量形式为基于秩的卡方统计量：

$$H_0 : \text{各组样本来自具有相同分布的总体} \quad \text{对比} \quad H_1 : \text{至少有一组的总体分布不同.} \tag{2}$$

SAS 在 `PROC NPAR1WAY` 的输出中同时给出了 Kruskal–Wallis 检验的卡方统计量和对应的 p 值。在结果汇报中，本报告将两种检验的结论一并呈现。

为直观展示 5 个指标的非参数检验结果，本报告将 5 张来自 SAS 的检验结果截图分别插入，对应 PTS、AST、TRB、BLK、STL 的 Wilcoxon 与 Kruskal–Wallis 检验输出，如图 6–10 所示。

Wilcoxon 评分（秩和） - 变量 PTS
按以下变量分类: Group

Group	N	评分和	H0 之下的期望值	H0 之下的标准差	均值评分
Frontcourt	183	27965.50	29005.50	801.815629	152.816940
Guard	133	22120.50	21080.50	801.815629	166.319549

平均评分用于结值。

Wilcoxon 双样本检验	
统计量	22120.5000
近似正态分布	
Z	1.2964
单侧 Pr > Z	0.0974
双侧 Pr > |Z|	0.1948
t 近似值	
单侧 Pr > Z	0.0979
双侧 Pr > |Z|	0.1958
Z 包括 0.5 的连续性校正。	

Kruskal-Wallis 检验	
卡方	1.6824
自由度	1
Pr > 卡方	0.1946

图 6: PTS 的 Wilcoxon 与 Kruskal–Wallis 检验结果示意（SAS 输出）

Wilcoxon 评分（秩和） - 变量 AST
按以下变量分类: Group

Group	N	评分和	H0 之下的期望值	H0 之下的标准差	均值评分
Frontcourt	183	23047.50	29005.50	801.550201	125.942623
Guard	133	27038.50	21080.50	801.550201	203.296992

平均评分用于结值。

Wilcoxon 双样本检验	
统计量	27038.5000
近似正态分布	
Z	7.4325
单侧 Pr > Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr > Z	<.0001
双侧 Pr > |Z|	<.0001

Z 包括 0.5 的连续性校正。

Kruskal-Wallis 检验	
卡方	55.2509
自由度	1
Pr > 卡方	<.0001

图 7: AST 的 Wilcoxon 与 Kruskal–Wallis 检验结果示意（SAS 输出）

Wilcoxon 评分（秩和） - 变量 TRB
按以下变量分类: Group

Group	N	评分和	H0 之下的期望值	H0 之下的标准差	均值评分
Frontcourt	183	35988.0	29005.50	801.674577	196.655738
Guard	133	14098.0	21080.50	801.674577	106.000000

平均评分用于结值。

Wilcoxon 双样本检验	
统计量	14098.0000
近似正态分布	
Z	-8.7093
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001

Z 包括 0.5 的连续性校正。

Kruskal-Wallis 检验	
卡方	75.8622
自由度	1
Pr > 卡方	<.0001

图 8: TRB 的 Wilcoxon 与 Kruskal–Wallis 检验结果示意（SAS 输出）

Wilcoxon 评分（秩和） - 变量 BLK
按以下变量分类: Group

Group	N	评分和	H0 之下的期望值	H0 之下的标准差	均值评分
Frontcourt	183	35459.0	29005.50	797.250472	193.765027
Guard	133	14627.0	21080.50	797.250472	109.977444

平均评分用于结值。

Wilcoxon 双样本检验	
统计量	14627.0000
近似正态分布	
Z	-8.0941
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001
t 近似值	
单侧 Pr < Z	<.0001
双侧 Pr > |Z|	<.0001

Z 包括 0.5 的连续性校正。

Kruskal-Wallis 检验	
卡方	65.5241
自由度	1
Pr > 卡方	<.0001

图 9: BLK 的 Wilcoxon 与 Kruskal–Wallis 检验结果示意（SAS 输出）

Wilcoxon 评分（秩和） - 变量 STL
按以下变量分类: Group

Group	N	评分和	H0 之下的期望值	H0 之下的标准差	均值评分
Frontcourt	183	26662.50	29005.50	797.854667	145.696721
Guard	133	23423.50	21080.50	797.854667	176.116541

平均评分用于结值。

Wilcoxon 双样本检验	
统计量	23423.5000
近似正态分布	
Z	2.9360
单侧 Pr > Z	0.0017
双侧 Pr > |Z|	0.0033
t 近似值	
单侧 Pr > Z	0.0018
双侧 Pr > |Z|	0.0036

Z 包括 0.5 的连续性校正。

Kruskal-Wallis 检验	
卡方	8.6238
自由度	1
Pr > 卡方	0.0033

图 10: STL 的 Wilcoxon 与 Kruskal–Wallis 检验结果示意（SAS 输出）

2.5 关键指标的分组比较结果

结合描述统计与两种非参数检验的结果，对 5 个关键指标的主要发现总结如下。数值为本次分析得到的近似值，用于说明结果趋势。

2.5.1 得分 PTS

1. 后卫的场均得分略高于前场，约为 13.11 分对 12.08 分。
2. Wilcoxon 秩和检验和 Kruskal–Wallis 检验的 p 值均约为 0.195，大于显著性水平 0.05。
3. 在显著性水平 0.05 下，不能拒绝“两个位置在得分分布上无显著差异”的原假设。

2.5.2 助攻 AST

1. 后卫的场均助攻明显高于前场，约为 3.86 次对 2.25 次。
2. Wilcoxon 秩和检验和 Kruskal–Wallis 检验的 p 值均约为 1.1×10^{-13}，远小于 0.01。
3. 在常用显著性水平下，可以认为两组在助攻分布上的差异极为显著。

2.5.3 篮板 TRB

1. 前场的场均篮板显著高于后卫，约为 5.54 个对 3.36 个。
2. Wilcoxon 秩和检验和 Kruskal–Wallis 检验的 p 值均约为 3.1×10^{-18}，远小于 0.01。
3. 说明前场与后卫在篮板能力上存在极显著差异。

2.5.4 盖帽 BLK

1. 前场的场均盖帽大约是后卫的 2 倍，约为 0.70 次对 0.34 次。
2. Wilcoxon 秩和检验和 Kruskal–Wallis 检验的 p 值均约为 5.8×10^{-16}，远小于 0.01。
3. 可以认为两组在盖帽能力上的差异极为显著。

2.5.5 抢断 STL

1. 后卫的场均抢断略高于前场，约为 0.88 次对 0.76 次。
2. Wilcoxon 秩和检验和 Kruskal–Wallis 检验的 p 值均约为 0.0033，小于 0.01。
3. 在常见显著性水平下，可以认为两组在抢断分布上的差异显著。

2.6 综合结果汇总表

为了直观展示上述比较结果，可以将 5 个指标的均值和显著性结论汇总于表 1 中。

表 1: 后卫与前场在 5 个关键指标上的比较概览

指标	后卫均值	前场均值	p 值范围	显著性结论
PTS	13.11	12.08	约为 0.20	差异不显著
AST	3.86	2.25	远小于 0.001	后卫显著更高
TRB	3.36	5.54	远小于 0.001	前场显著更高
BLK	0.34	0.70	远小于 0.001	前场显著更高
STL	0.88	0.76	小于 0.01	后卫显著更高

3 所得结论

基于 2023–2024 赛季 NBA 常规赛主要轮换球员的数据，本次课程设计围绕球员位置与 5 个关键技术指标之间的关系进行了系统分析，主要结论概括如下。

1. 得分层面。在满足出场数和上场时间条件的轮换球员样本中，后卫的场均得分略高于前场，但 Wilcoxon 与 Kruskal–Wallis 检验给出的 p 值均约为 0.195，差异未达到显著水平。说明在现代篮球体系下，得分责任更加均衡，前场球员在进攻端的作用已经显著提升。

2. 组织进攻层面。后卫在场均助攻上显著高于前场，差异达到极高显著性，与后卫承担主要持球和组织任务的传统角色定位高度一致，说明助攻是区分后卫与前场功能的重要指标。

3. 篮板与护筐层面。前场球员在场均篮板和盖帽上均显著领先后卫，且差异幅度较大。这体现了前场在篮板保护、内线防守和护筐方面的核心作用，与其身材条件和站位特点相吻合。

4. 外线防守与抢断层面。后卫在场均抢断上显著高于前场，体现出后卫在外线防守中更加积极的抢断和预判。这与后卫频繁负责持球人防守的实际情况相符。

5. 统计方法层面。通过对 5 个指标进行描述统计、可视化，并结合 Wilcoxon 秩和检验与 Kruskal–Wallis 检验，本次分析证明了非参数检验在处理非正态分布、样本方差不齐时的适用性，为今后在体育数据及其他领域中应用统计方法提供了参考。

总体而言，本次课程设计从球员位置这一直观维度出发，利用较为完整的赛季统计数据，验证并量化了篮球战术常识中关于角色分工的若干判断，也展示了统计分析在体育数据挖掘中的实际价值。

4 遇到的困难及解决方法

在课程设计的完成过程中，小组遇到了一些具体问题和困难，主要集中在数据理解、分析思路和工具使用 3 个方面。下面对这些问题及相应解决方法进行总结。

4.1 原始数据量大、变量众多

问题描述 初始数据集包含 700 余名球员和 30 多个技术指标，如果不加筛选地全部纳入分析，不仅计算量较大，而且可能被大量出场不稳定的边缘球员噪声干扰，难以得出稳健结论。

解决方法 小组查阅课程资料并与老师交流，最终确立了“出场数不少于 20 场、场均时间不少于 15 分钟”的筛选标准，以保证样本的稳定性与代表性；同时根据研究问题，从众多指标中筛选出 PTS、AST、TRB、BLK、STL 5 个关键技术指标，在保证分析聚焦性的同时仍能较全面地反映进攻、防守和效率特征。

4.2 分析维度选择与思路调整

问题描述 在项目初期，小组曾考虑从“单个球员纵向表现”或者“球队整体表现预测”等角度展开分析，但由于本次数据仅为赛季平均值，且不同球队在筛选后样本数和出场时间差异较大，直接做球队层面的比较并不合适。

解决方法 经过讨论，小组决定将分析重点转移到球员位置维度，按照后卫与前场进行分组，比较不同位置在关键指标上的分布差异。这一思路既避免了球队层面样本不均衡的问题，又能紧密结合篮球战术逻辑，方便对统计结果进行解释。

4.3 主要技术指标含义理解不足

问题描述 原始数据包含大量缩写指标，例如 PTS、AST、TRB、STL、BLK、eFG% 等，部分组员对其具体含义和统计口径并不熟悉，在变量筛选和结果解读时容易产生混淆。

解决方法 小组查阅 NBA 官方统计说明和相关资料，对每一个候选指标的含义、计算方式和实际意义进行梳理，并在组内统一口径。只有在充分理解指标含义之后，才将其纳入正式分析，从而保证结论的可靠性与可解释性。

4.4 SAS 代码与语法掌握不熟练

问题描述 部分组员在使用 SAS 进行数据导入、描述统计、制图以及非参数检验时，一开始对过程步骤和 ODS 输出来说并不熟悉，容易出现变量名不合法、路径设置错误或图形输出失败等问题。

解决方法 小组通过以下几种方式解决相关问题。

1. 查阅 SAS 帮助文档和网络资料，了解 `PROC IMPORT`、`PROC MEANS`、`PROC UNIVARIATE`、`PROC SGPLOT`、`PROC NPAR1WAY` 等过程的基本用法。
2. 参考示例程序，先在小规模数据上进行测试，再逐步移植到本项目中。
3. 通过 ODS 将结果导出为 HTML 或 Excel 文件，便于检查输出是否符合预期。

通过多次调试，最终搭建起一套相对完整且可复用的 SAS 分析代码，不仅满足了本次课程设计的需求，也为今后进行类似统计分析打下了基础。

4.5 非参数检验结果的理解与解释

问题描述 Wilcoxon 秩和检验和 Kruskal–Wallis 检验属于非参数方法，与更为常见的 t 检验相比，其统计量形式与 p 值的含义不够直观，在解释“差异显著或不显著”时需要格外谨慎。

解决方法 小组在查阅教材和相关资料的基础上，从秩的角度重新理解了非参数检验的原理，并结合箱线图、直方图等可视化结果进行交叉验证。当箱线图中两组中位数和分布区间差异较大时，通常对应较小的 p 值；反之则对应较大的 p 值。通过这种“图形与数值结合”的方式，既避免了机械套用公式，也提高了对结果的直观理解能力。

参考资料

1. NBA 官方技术统计说明以及 2023–2024 赛季球员数据网站。
2.《数据统计分析》课程教材。

3. 小组整理的 SAS 程序与导出结果。