清华大学数据库技术与应用

数据统计 I

授课教师：计算机系王健楠

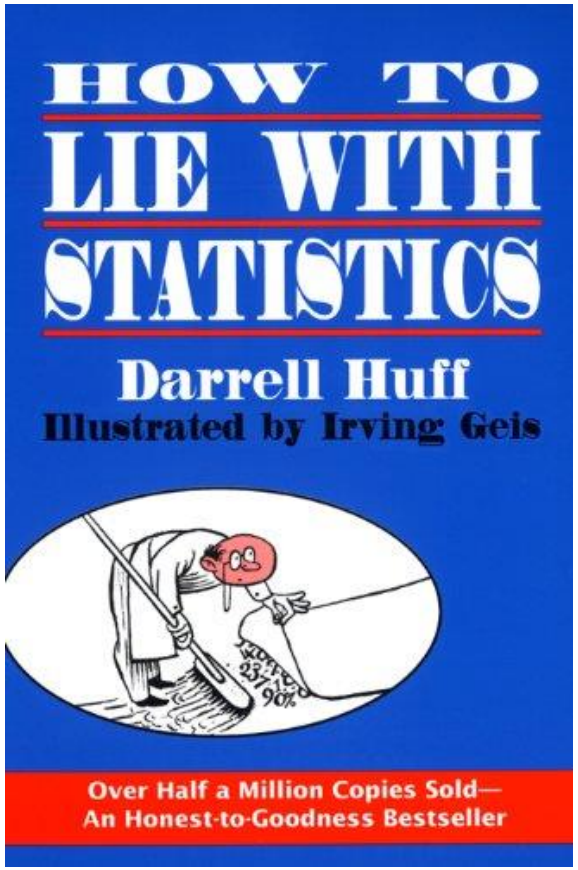
授课学期：2026年（春季）

清華大學 Tsinghua University

为什么要学统计学?

统计学家 Mark Twain 曾说: ***"There are three kinds of lies: lies, damned lies, and statistics"***

统计数据无处不在，但经常被误用或误读。本课帮助你建立正确的统计思维，避免被统计数据误导。

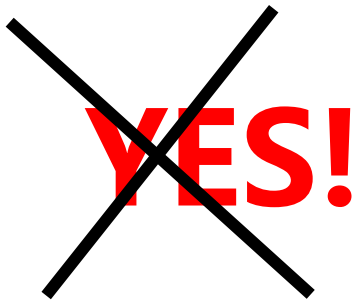


1. The Sample with the Built-in Bias 13
2. The Well-Chosen Average 29
3. The Little Figures That Are Not There 39
4. Much Ado about Practically Nothing 55
5. The Gee-Whiz Graph 62
6. The One-Dimensional Picture 68
7. The Semiattached Figure 76
8. Post Hoc Rides Again 89
9. How to Statisticulate 102
10. How to Talk Back to a Statistic 124

辛普森悖论（Simpson's Paradox）

问题：UC Berkeley 在研究生录取中存在性别偏见吗?

	Applicants	Admitted
Men	8442	44%
Women	4321	35%



辛普森悖论（Simpson's Paradox）

问题：UC Berkeley 在研究生录取中存在性别偏见吗?

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

NO!

- 女性倾向于报名竞争激烈、录取率低的热门院系
- 在同一院系内，录取率没有显著的性别差异

课程大纲

清華大學 Tsinghua University

01
统计思维

02
描述性统计

03
推断性统计

课程大纲

清華大學 Tsinghua University

01 统计思维

- 样本、总体及其关系
- 有无统计思维的对比

02 描述性统计

03 推断性统计

统计思维（Statistical Thinking）

清華大學 Tsinghua University

- 数据只是样本
- 你的目标是：从样本推断总体
- 思考如何"逆向推理"：从样本回溯到总体

例1：图像分类

这张图片是猫还是狗?

数据集

- 1000 张从网络收集的图片
- 包含猫和狗的标注标签

没有统计思维

将这 1000 张图片视为“总体”本身：

- 在全部数据上训练模型
- 在相同的数据上评估模型
- 得到结论：模型准确率 = 95%

具有统计思维的做法

清華大學 Tsinghua University

什么是总体?

- 网络上的所有图片

什么是你的数据集?

- 从网络上随机抽取的 1000 张图片（样本）

应该怎么做?

- 将数据集划分为训练集和测试集
  - 例如：80% 训练，20% 测试
- 在训练集上训练模型
- 在测试集上评估模型

例2：民调预测

谁将赢得这次选举？

TRUMP vs. Hillary

- 对 1000 名选民进行调查
- 记录他们的投票意向

没有统计思维

将这 1000 人视为"总体"本身：

- 统计希拉里的支持者数量，例如：520 人
- 统计特朗普的支持者数量，例如：480 人
- 得出结论：希拉里将赢得选举

具有统计思维的做法

什么是总体?

- 所有将在选举日参与投票的人

什么是你的数据集?

- 选举前对 1000 人进行的调查（样本）

分析结果（含误差范围）

- 希拉里：52% ± 3%
- 特朗普：48% ± 2%

重要假设：样本中的人，从调查时到选举日之间没有改变投票意向。

课程大纲

01
统计思维

02
描述性统计

- 描述性统计 vs. 推断性统计
- **Task-Centric EDA**
- 相关性分析

03
推断性统计

描述性统计 vs. 推断性统计

清華大學 Tsinghua University

描述性统计

（例：中位数 Median）

目的 (Why)

- 理解数据本身

方法 (How)

- 数据汇总（均值、中位数、方差等）
- 数据可视化

推断性统计

（例：A/B 测试）

目的 (Why)

- 用样本数据推断总体规律

方法 (How)

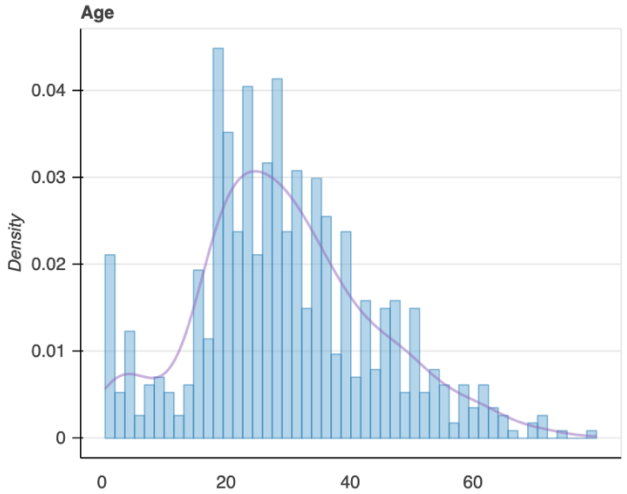
- 参数估计 (Estimation)
- 置信区间 (Confidence Intervals)
- 假设检验 (Hypothesis Testing)

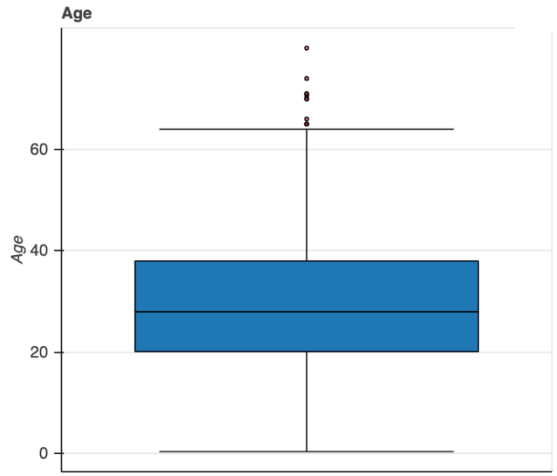
探索性数据分析（EDA）

通过数据可视化、数据汇总等方式，理解数据并发现洞察。

理解"年龄"列（Age 列）

Minimum	0.42
5-th Percentile	4
Q1	20.125
Median	28
Q3	38
95-th Percentile	56
Maximum	80
Range	79.58
IQR	17.875





Python 中的 EDA 方案（当前现状）

方案 1：Pandas + Matplotlib

☹难以使用

- 初学者：需要学会编写绘图代码
- 专家：需要编写大量重复性代码

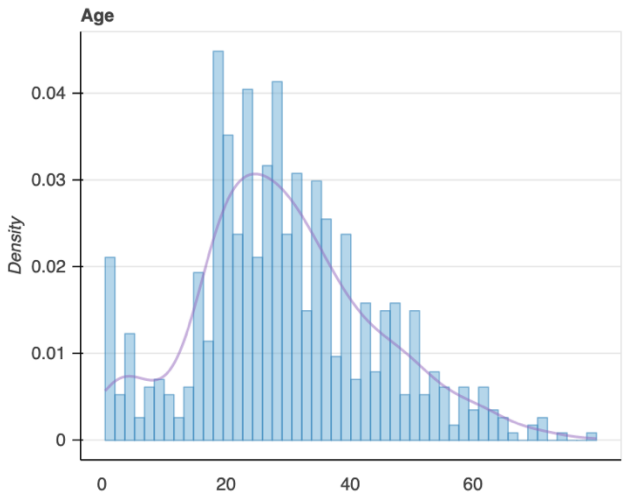
理解"年龄"列（Age 列）

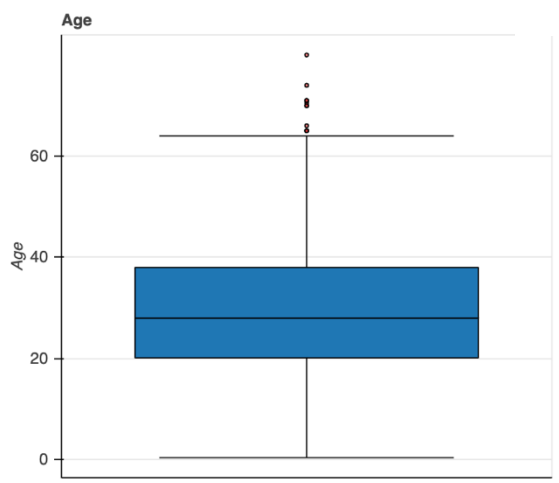
写代码

写代码

写代码

Minimum	0.42
5-th Percentile	4
Q1	20.125
Median	28
Q3	38
95-th Percentile	56
Maximum	80
Range	79.58
IQR	17.875





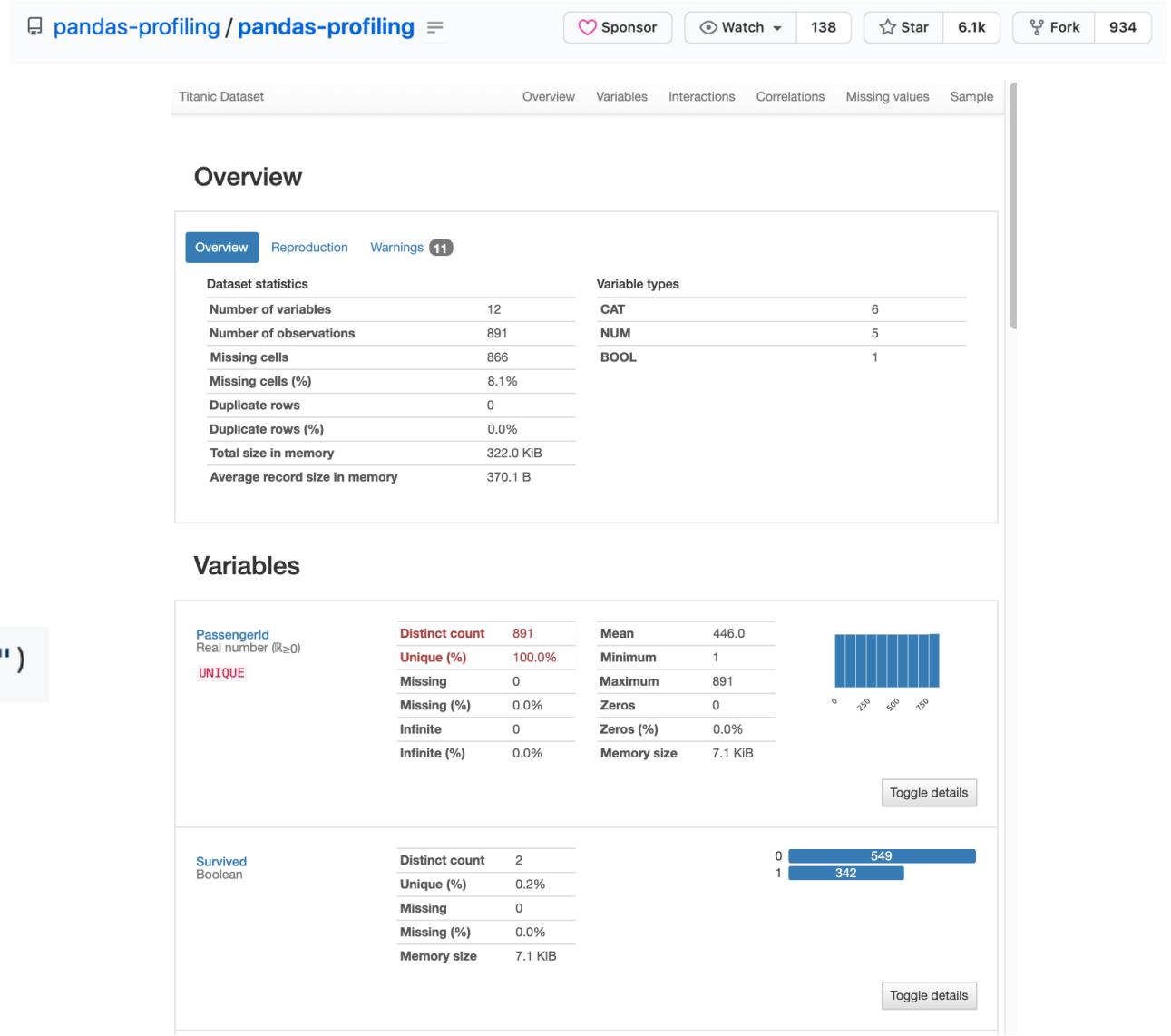
Python 中的 EDA 方案（当前现状）

方案 2：Pandas-profiling

☹ 速度慢

☹ 难以定制

```
profile = ProfileReport(df, title="Pandas Profiling Report")
```



DataPrep.EDA 设计目标

清華大學 Tsinghua University

EDA 方案	易用性	交互速度	易定制性
1. Pandas + Matplotlib	☹	☺	☺
2. Pandas-profiling	☺	☹	☹
3. DataPrep.EDA	☺	☺	☺

课程大纲

01
统计思维

02
描述性统计

- 描述性统计 vs. 推断性统计
- Task-Centric EDA
- **相关性分析**

03
推断性统计

核心设计理念：Task-Centric EDA

清華大學 Tsinghua University

以任务为中心的 API 设计

- 声明式（Declarative）
- 同时支持粗粒度和细粒度的 EDA 任务

API 示例：

- **plot(df)** → "我想查看数据集的整体概览"
- **plot_missing(df)** → "我想了解数据集中的缺失值情况"
- **plot(df, x)** → "我想理解列 x 的分布"
- **plot(df, x, y)** → "我想理解 x 和 y 之间的关系"

相关性分析（Correlation Analysis）

清華大學 Tsinghua University

什么是相关性（Correlation）？

- 相关性是衡量两个变量之间关系的度量。

为什么相关性分析有用?

- 更好地理解数据：揭示变量之间的隐藏关系
- 提升预测效果：找到与目标变量相关的特征

案例分析：如何进行相关性分析

清華大學 Tsinghua University

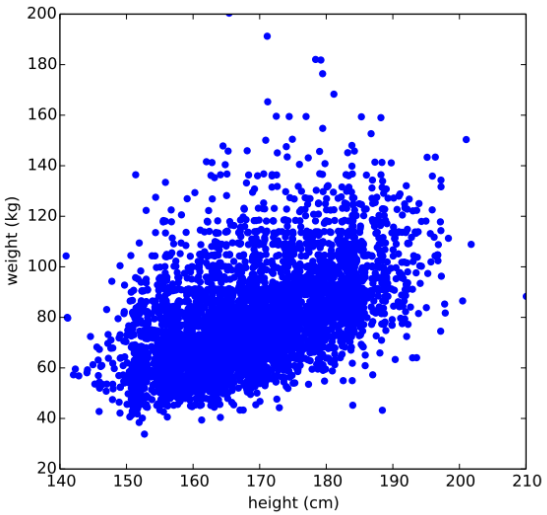
案例：身高与体重的相关性

	height	weight	age	male
1				
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0

方法 1：可视化

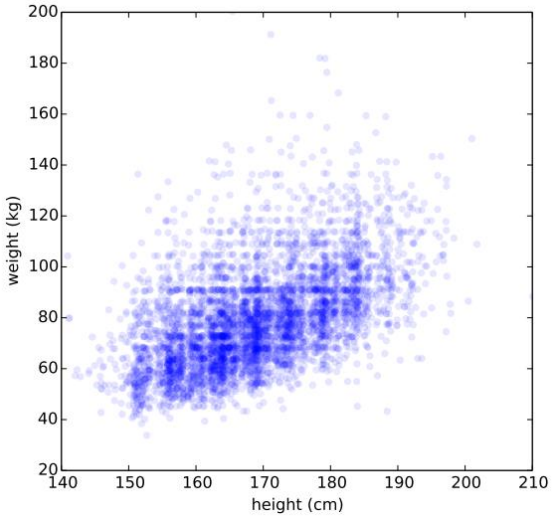
散点图

	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



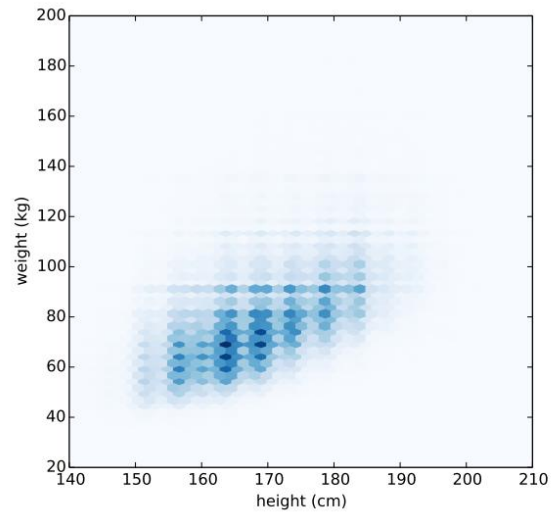
散点图（含透明度）

	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



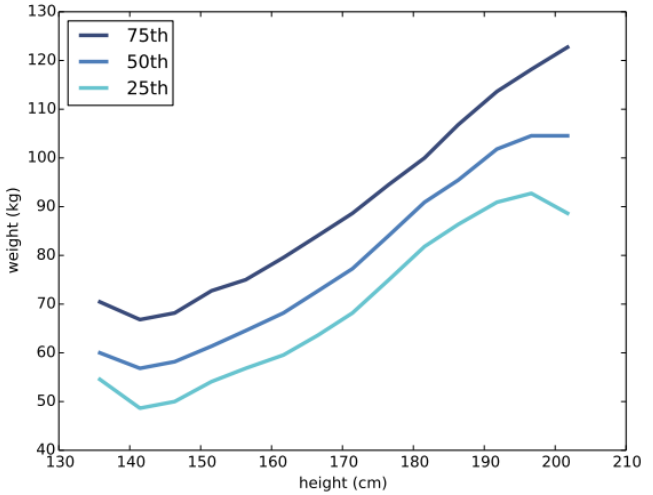
六边形密度图（Hexbin Plot）

	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



描述变量关系的特征

	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



方法 2：相关系数

用数值定量描述两个变量之间的线性关系强度与方向

协方差（Covariance）

协方差（Covariance）：衡量两个变量同向变化趋势的统计量。

$$\operatorname{cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right]$$

$$\operatorname{cov}(X,Y) = \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y]$$

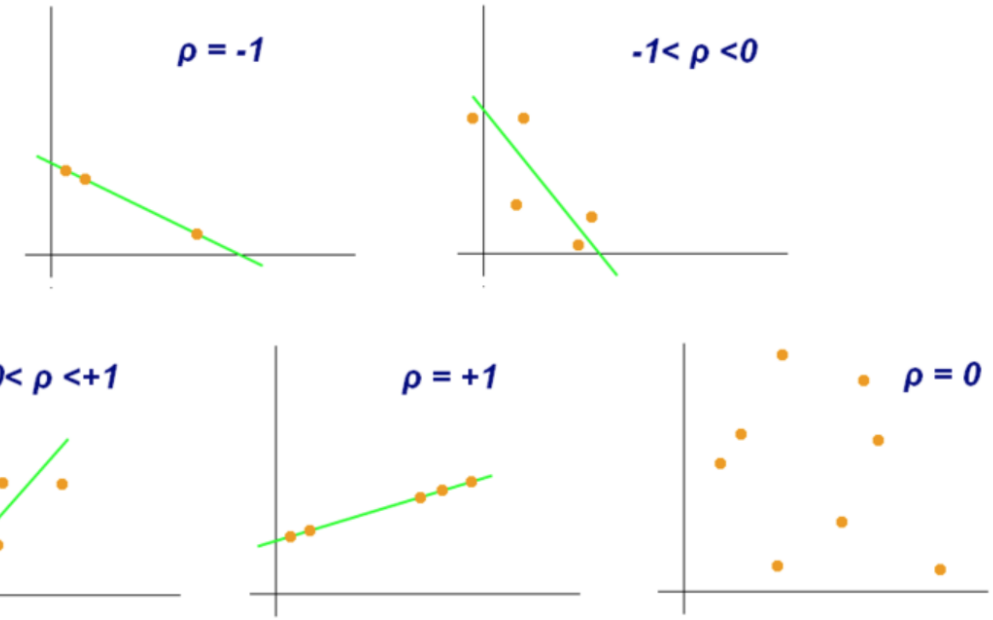
局限性：难以解读

- 结果带有单位（如：千克·厘米），难以直接比较
- 示例：结果为 113 千克·厘米 → 这代表相关性强还是弱？无法判断

皮尔逊相关系数（Pearson's Correlation）

皮尔逊相关系数：衡量两个变量之间线性关系的标准化度量。

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$



易于解读（范围 -1 到 +1）

- [-1, 0)：负相关
- (0, +1]：正相关
- -1 或 +1：完全线性相关

局限性

- 只能衡量线性关系
- 对异常值非常敏感

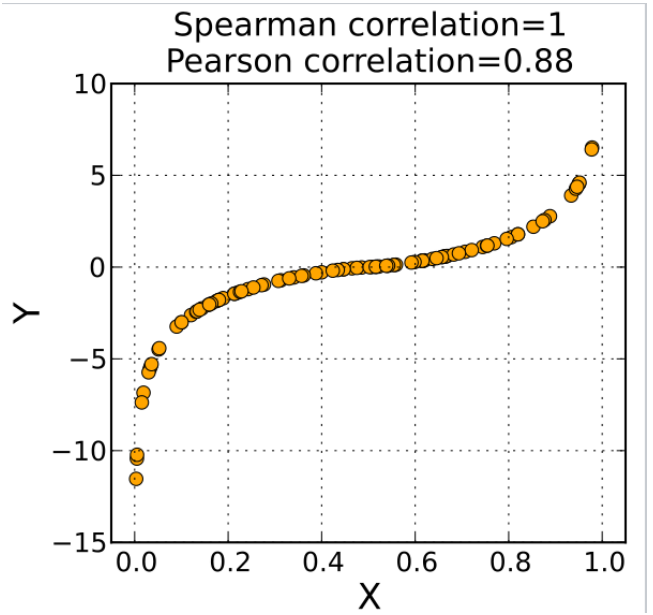
非线性关系呢?

斯皮尔曼秩相关系数 (Spearman's rank correlation)

斯皮尔曼秩相关系数：衡量两个变量之间单调关系的度量。

$$r_s = \rho_{\mathrm{r}_X, \mathrm{r}_Y} = \frac{\operatorname{cov}(\mathrm{r}_X, \mathrm{r}_Y)}{\sigma_{\mathrm{r}_X} \sigma_{\mathrm{r}_Y}}$$

核心思想：先将原始数据转换为排名（Rank），再计算皮尔逊相关系数。



优势

- 对异常值更加鲁棒
- 对Skewed分布更稳健
- 可处理非线性的单调关系

课堂练习：相关性分析

以下是5名学生的学习时间与考试成绩数据：

学生编号	学习时间 X (小时)	考试成绩 Y (分)
1	1	2
2	2	4
3	3	5
4	4	4
5	5	8

请根据左侧数据完成以下计算：

1. 计算协方差 Cov(X, Y)

$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

2. 计算皮尔逊相关系数 r

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

3. 计算斯皮尔曼秩相关系数 ρ

$$r_s = \rho_{\mathrm{r}_X,\mathrm{r}_Y} = \frac{\operatorname{cov}(\mathrm{r}_X,\mathrm{r}_Y)}{\sigma_{\mathrm{r}_X}\sigma_{\mathrm{r}_Y}}$$

课堂练习：详细解答

1 协方差 (Covariance)

Cov = 2.40

$Cov(X,Y) = \Sigma[(X-\bar{X})(Y-\bar{Y})] / N$

= 12.0 / 5 = **2.40**

2 皮尔逊相关系数 (Pearson's r)

r = 0.87

$r = Cov(X,Y) / (\sigma_X \times \sigma_Y)$

= 2.40 / (1.41 × 1.96)
= 2.40 / 2.76 ≈ **0.87**

结论：学习时间与成绩存在强正相关。

3 斯皮尔曼秩相关系数 Ã ÕÑMØÒ MÒÃCEtiÅ

ρ = 0.82

排名转换 (注意Y=4出现两次，取平均排名2.5)

X值	1	2	3	4	5
Rank X	1	2	3	4	5
Y值	2	4	5	4	8
Rank Y	1	2.5	4	2.5	5

Cov(Rank) = 1.60

σ_RX = 1.41, σ_RY = 1.38

ρ = 1.60 / (1.41×1.38) ≈ **0.82**

本讲小结

统计思维

- 样本与总体的关系
- 有无统计思维的对比分析

描述性统计

- 描述性统计 vs. 推断性统计
- Task-Centric EDA
- 相关性分析（协方差、皮尔逊、斯皮尔曼）

推断性统计

- （待续）