

• Statistics 3: Wilcoxon signed-rank test

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed.

The test is named for Frank Wilcoxon (1892–1965) who, in a single paper, proposed both it and the rank-sum test for two independent samples (Wilcoxon, 1945). The test was popularized by Siegel (1956) in his influential text book on non-parametric statistics.

			$x_{2,i} - x_{1,i}$	
i	$x_{2,i}$	$x_{1,i}$	sgn	abs
1	125	110	1	15
2	115	122	-1	7
3	130	125	1	5
4	140	120	1	20
5	140	140		0
6	115	124	-1	9
7	140	123	1	17
8	125	137	-1	12
9	140	135	1	5
10	135	145	-1	10

order by absolute difference

			$x_{2,i} - x_{1,i}$			
i	$x_{2,i}$	$x_{1,i}$	sgn	abs	R_i	sgn · R_i
5	140	140		0		
3	130	125	1	5	1.5	1.5
9	140	135	1	5	1.5	1.5
2	115	122	-1	7	3	-3
6	115	124	-1	9	4	-4
10	135	145	-1	10	5	-5
8	125	137	-1	12	6	-6
1	125	110	1	15	7	7
7	140	123	1	17	8	8
4	140	120	1	20	9	9

$$N_r = 10 - 1 = 9, W = |1.5 + 1.5 - 3 - 4 - 5 - 6 + 7 + 8 + 9| = 9.$$

$$W < W_{\alpha=0.05,9} = 35, \therefore \text{fail to reject } H_0$$



LIS scholars involving computer science research

$R(h_WoS) > R(h_GS)$ ($\rho < 0.05$); $R(h_WoS) > R(h_SCO)$ ($\rho < 0.01$)

AUTHOR	R(h_GS)	R(h_SCO)	R(h_WoS)
Goker, Ayse	96	100	100
Jarvelin, Kalervo	24	15	26
Jose, Joemon M	72	67	76
Kantor, Paul B	41	36	62
Lalmas, Mounia	17	43	62
Liddy, Elizabeth DuRoss	37	81	76
Losee, Robert M	51	47	33
Ounis, Iadh	55	72	92
van Rijsbergen, CJ	6	15	76
Robertson, Stephen	81	30	62
Ruger, Stefan	87	72	76
Ruthven, Ian	41	47	62
Sanderson, Mark	24	47	76
Tait, John I.	77	88	97
Whittaker, Steve J.	3	15	87
Willett, Peter	1	1	1
Yang, Christoph C	72	30	62

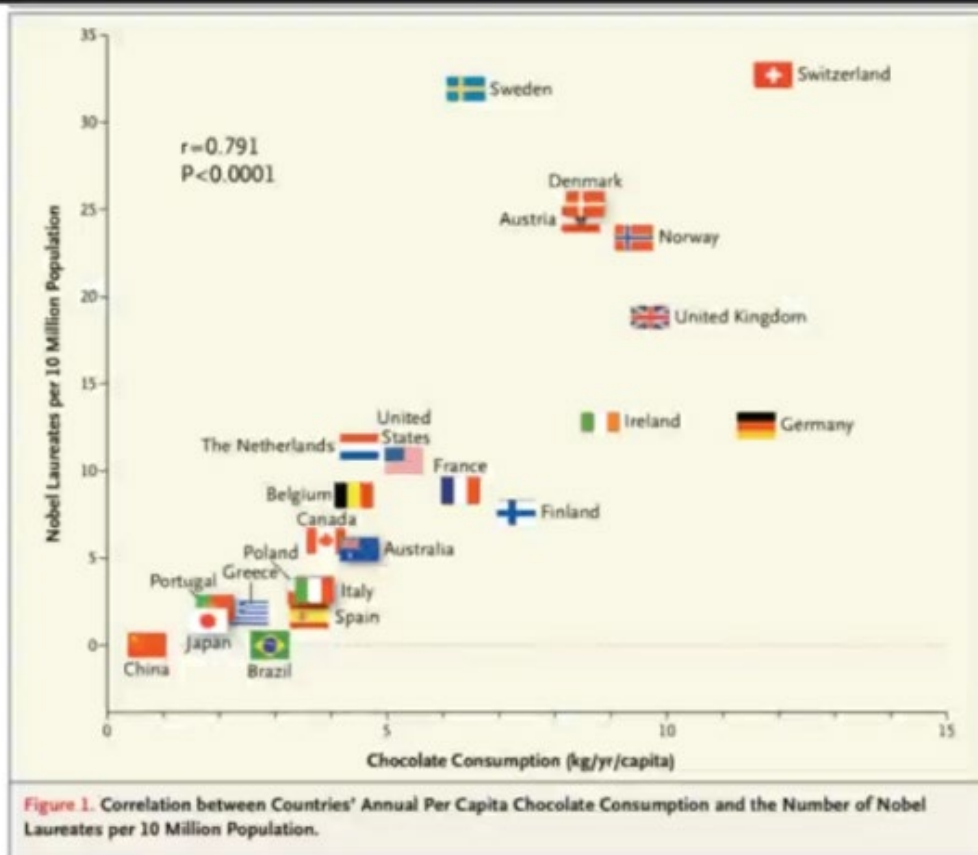
			$x_{2,i} - x_{1,i}$	
i	$x_{2,i}$	$x_{1,i}$	sgn	abs
1	125	110	1	15
2	115	122	-1	7
3	130	125	1	5
4	140	120	1	20
5	140	140		0
6	115	124	-1	9
7	140	123	1	17
8	125	137	-1	12
9	140	135	1	5
10	135	145	-1	10

order by absolute difference

			$x_{2,i} - x_{1,i}$			
i	$x_{2,i}$	$x_{1,i}$	sgn	abs	R_i	sgn · R_i
5	140	140		0		
3	130	125	1	5	1.5	1.5
9	140	135	1	5	1.5	1.5
2	115	122	-1	7	3	-3
6	115	124	-1	9	4	-4
10	135	145	-1	10	5	-5
8	125	137	-1	12	6	-6
1	125	110	1	15	7	7
7	140	123	1	17	8	8
4	140	120	1	20	9	9

$$N_r = 10 - 1 = 9, W = |1.5 + 1.5 - 3 - 4 - 5 - 6 + 7 + 8 + 9| = 9.$$

⇒ $W < W_{\alpha=0.05,9} = 35$, fail to reject H_0



四、 大数据时代：从知识回到数据

● 大数据是什么？

- 是一种大量而复杂的数据集合数据集合
- 在可承受的范围内，无法用传统数据库系统和常规软件和常规软件工具对内容进行获取、存储、管理和分析

● 大数据的特征 (5V)

- 容量 (Volume) : 数据量巨大
- 种类 (Variety) : 数据类型复杂多样
- 速度 (Velocity) : 快速甚至实时地采集、处理数据并做出正确反馈
- 价值 (Value) : 价值密度低
- 真实性 (Veracity) : 数据判断准确可靠

● 大数据的发展历程

- 20世纪末是大数据的萌芽期，处于数据挖掘技术阶段。一些商业智能工具和知识管理技术开始被应用。
- 社交网络的流行导致大量非结构化数据出现，传统处理方法难以应对，数据处理系统、数据库架构开始重新思考。
- 2006年-2009年，大数据形成并行计算和分布式系统，为大数据发展的成熟期。
- 2008年9月，《自然》杂志出版“big data”专刊，使“大数据”这一概念在学术界得到认可和广泛使用

● 大数据的发展历程

- 2010年以来，随着智能手机应用，数据碎片化、分布式、流媒体特征更加明显。移动数据急剧增长。
- 2012年维克托·舍恩伯格《大数据时代：生活、工作与思维的大变革》宣传推广，大数据概念开始风靡全球
- 2013年5月，《颠覆性技术：技术改进生活、商业和全球经济》的研究报告确认了未来12种新兴技术，而大数据是这其中需求技术的基石
- 2014年5月，美国白宫发布的2014年全球“大数据”白皮书的研究报告《大数据：抓住机遇，守护价值》鼓励使用数据推动社会进步

● 大数据的发展趋势

➤ 数据的资源化

指大数据成为企业和社会关注的重要战略资源，并已成为大家争相抢夺的新焦点。因而，企业必须要提前制定大数据营销战略计划，抢占市场先机。

➤ 与云计算的深度结合

大数据离不开云处理，云处理为大数据提供了弹性可拓展的基础设备，是产生大数据的平台之一。

➤ 科学理论的突破

随着大数据的快速发展，其很有可能是新一轮的技术革命。随之兴起的数据挖掘、机器学习和人工智能等相关技术，可能会改变数据世界里的很多算法和基础理论，实现科学技术上的突破。

➤ 数据科学和数据联盟的成立

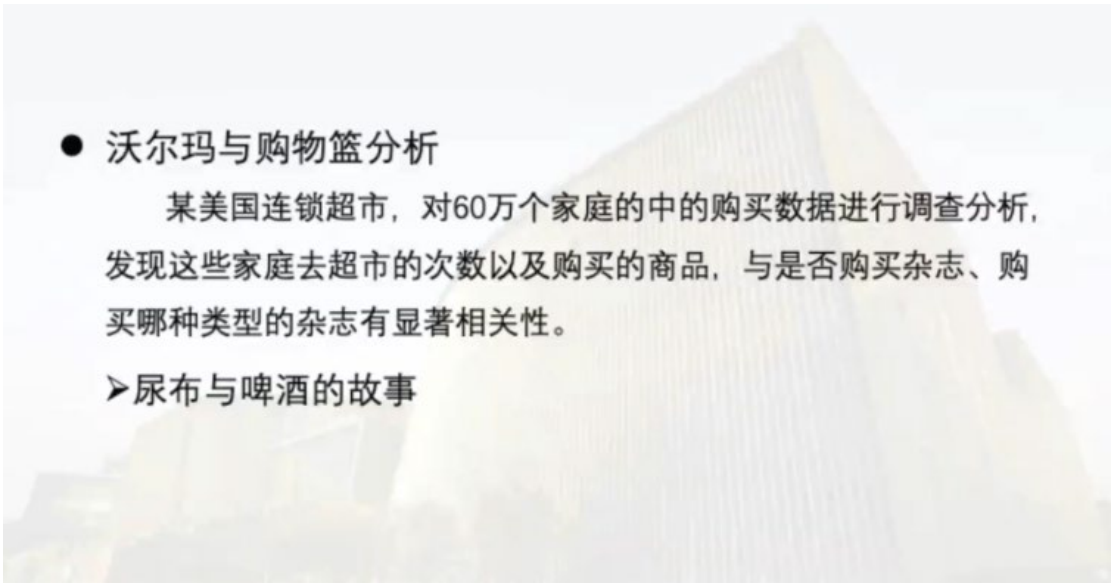
未来，数据科学将成为一门专门的学科，被越来越多的人所认知。

➤ 数据泄露泛滥

未来几年数据泄露事件的增长率也许会达到100%，除非数据在其源头就能够得到安全保障。企业需要从新的角度来确保自身以及客户数据，所有数据在创建之初便需要获得安全保障，而并非在数据保存的最后一个环节，仅仅加强后者的安全措施已被证明于事无补。

➤ 数据管理成为核心竞争力

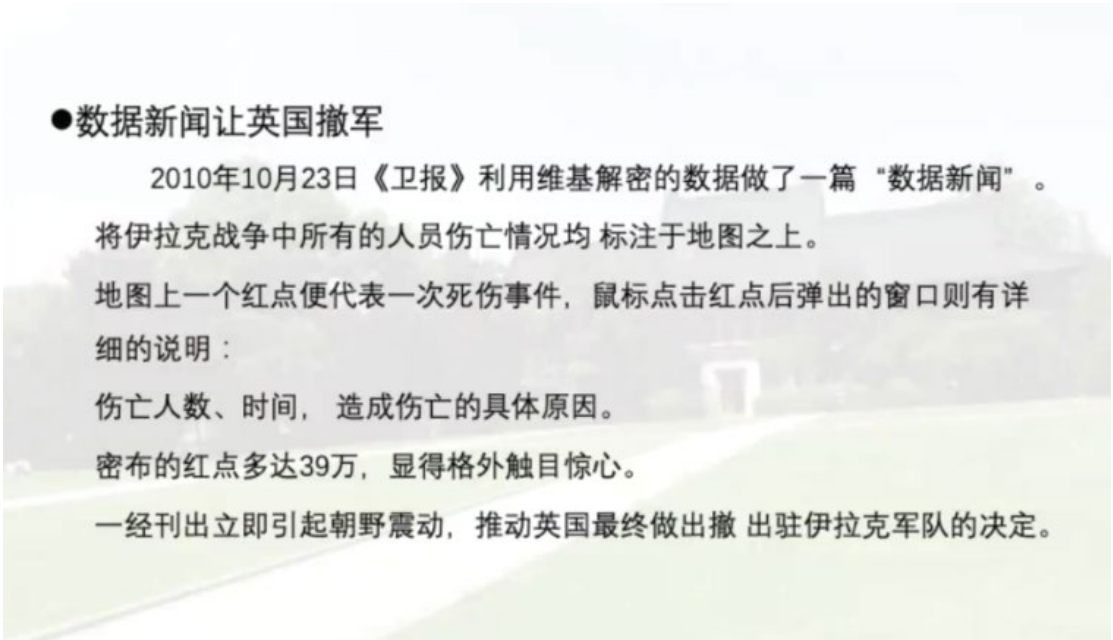
数据管理成为核心竞争力，直接影响财务表现。当“数据资产是企业核心资产”的概念深入人心之后，企业对于数据管理便有了更清晰的界定，将数据管理作为企业核心竞争力，持续发展，战略性规划与运用数据资产，成为企业数据管理的核心。数据资产管理效率与主营业务收入增长率、销售收入增长率显著正相关；此外，对于具有互联网思维的企业而言，数据资产竞争力所占比重为36.8%，数据资产的管理效果将直接影响企业的财务表现。



● 沃尔玛与购物篮分析

某美国连锁超市，对60万个家庭的中的购买数据进行调查分析，发现这些家庭去超市的次数以及购买的商品，与是否购买杂志、购买哪种类型的杂志有显著相关性。

➤ 尿布与啤酒的故事



● 数据新闻让英国撤军

2010年10月23日《卫报》利用维基解密的数据做了一篇“数据新闻”。将伊拉克战争中所有的人员伤亡情况均标注于地图之上。

地图上一个红点便代表一次死伤事件，鼠标点击红点后弹出的窗口则有详细的说明：

伤亡人数、时间，造成伤亡的具体原因。

密布的红点多达39万，显得格外触目惊心。

一经刊出立即引起朝野震动，推动英国最终做出撤出驻伊拉克军队的决定。

➤高德地图

众多用户通过手机实时反馈车速，地图 APP 预测道路拥堵



思考与讨论1：

信息社会中，为什么决策的依据从知识回到数据？

思考与讨论2：

大数据的终点在哪里？