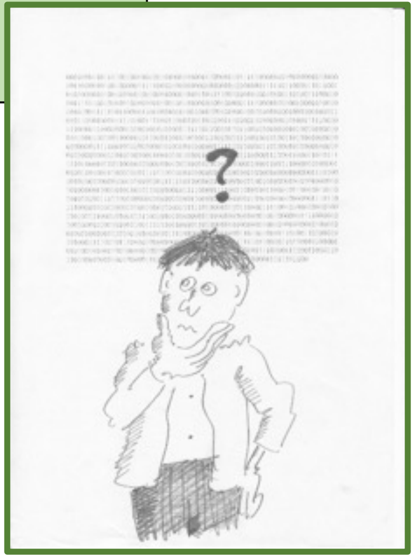
Correlations Refresher and Kendall Correlation

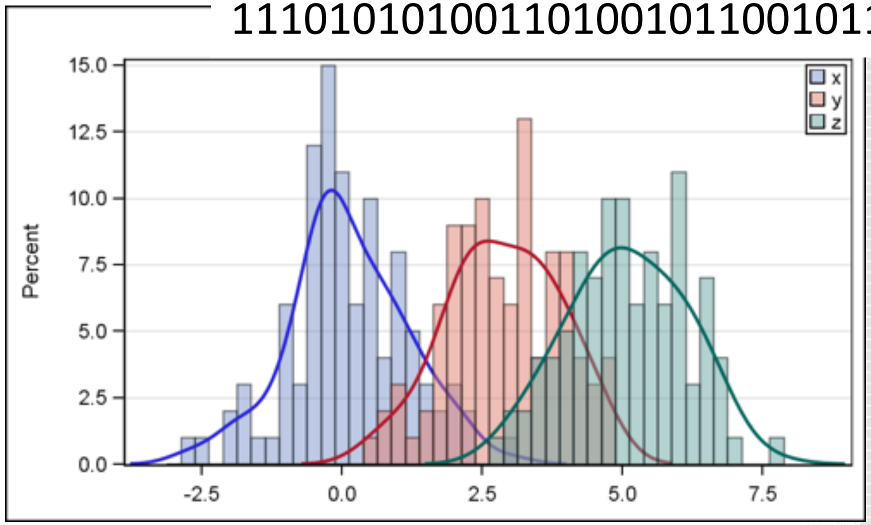
Statistics and data analysis

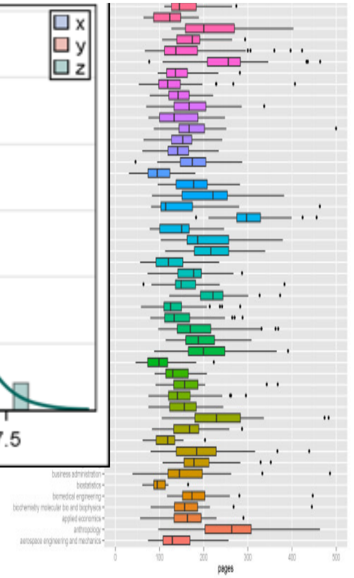
Zohar Yakhini

IDC, Herzeliya

001001110101010010101010010010001010101000101011111010110100110010011110101010011010010110010110110010







Pearson correlation

Population:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)V(Y)}}$$

Sample, for a particular realization $(\boldsymbol{x}, \boldsymbol{y})$ of n repeated sampling from (X,Y)

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n}(x_i - \mu(\boldsymbol{x}))(y_i - \mu(\boldsymbol{y}))}{\sqrt{\left(\sum_{i=1}^{n}\left(x_i - \mu(\boldsymbol{x})\right)^2\right)\left(\sum_{i=1}^{n}\left(y_i - \mu(\boldsymbol{y})\right)^2\right)}}$$

The Fisher Transform

$$F(r) = 0.5\ \ln\frac{1+r}{1-r}$$

Ronald Fisher
1890-1962

Thm (Fisher 1921):
If we start with (X, Y) that are close to bivariate normal then $F(\widehat{\rho_n})$, for i.i.d sampling, is normally distributed with mean $F\left(\rho = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}\right)$ and a standard deviation of $\frac{1}{\sqrt{n-3}}$.

Spearman's Rank Correlation Coefficient

Pearson:

$$\rho(x, y) = \frac{\sum_{i=1}^{n}(x_i - \mu(\boldsymbol{x}))(y_i - \mu(\boldsymbol{y}))}{\sqrt{\left(\sum_{i=1}^{n}\left(x_i - \mu(\boldsymbol{x})\right)^2\right)\left(\sum_{i=1}^{n}\left(y_i - \mu(\boldsymbol{y})\right)^2\right)}}$$

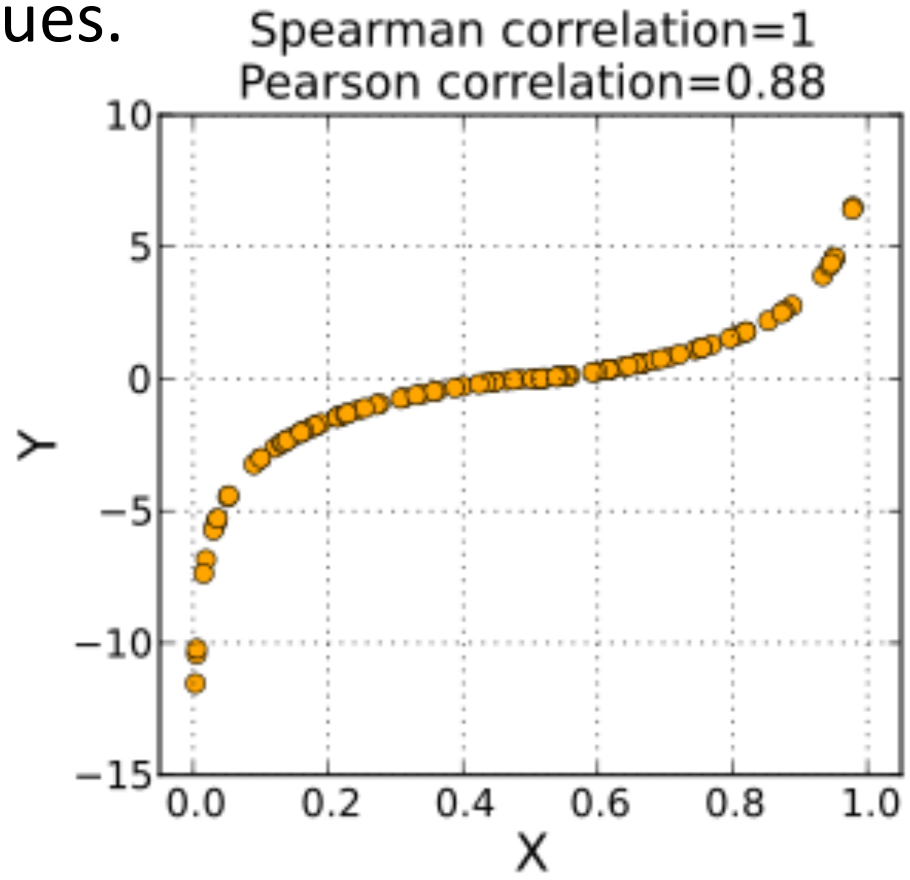
Spearman:

$$\mathrm{SP}(x, y) = \frac{\sum_{i=1}^{n}(u_i - \frac{n+1}{2})(v_i - \frac{n+1}{2})}{\sum_{i=1}^{n}\left(u_i - \frac{n+1}{2}\right)^2}$$

$$\boxed{\begin{aligned} u_i &= rank(x_i) \\ v_i &= rank(y_i) \end{aligned}}$$

Spearman rank correlation

- Perform Pearson correlation on the rank values.
- Ties can be handled by fractional ranks.
- $-1 \le SRC \le 1$, always …
- When is it -1? 1?



Spearman p-values

Let σ, π be uniformly drawn permutations in S_n.

Let $\rho = \rho(\sigma, \pi)$ be their Spearman correlation.

If σ, π are independently drawn then

$$Z = F(\rho)\sqrt{\frac{n-3}{1.06}} \sim N(0,1)$$

where F is the Fisher Transform: $F(r) = \frac{1}{2}\ln\frac{1+r}{1-r}$

Kendall correlation coefficient

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be the observed data.
Assume that all values of (x_i) and (y_i) are unique.
A pair of observations (x_i, y_i) and (x_j, y_j) is said to be ***concordant*** if the rank orders for both coordinates agree.

$x_i > x_j$ and $y_i > y_j$ OR $x_i < x_j$ and $y_i < y_j$.

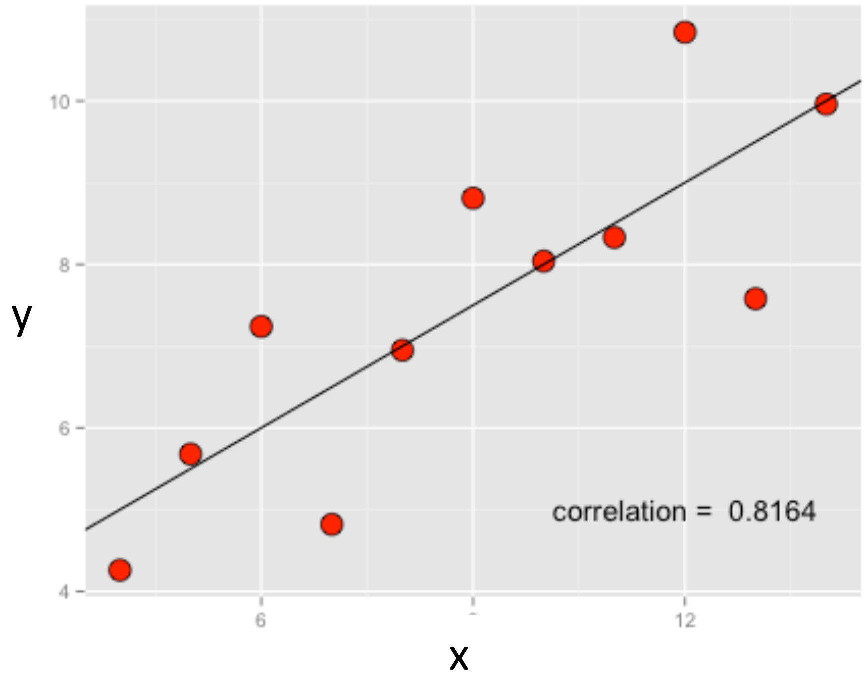
It is said to be ***discordant*** if

$x_i > x_j$ and $y_i < y_j$ OR $x_i < x_j$ and $y_i > y_j$.

If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant. (and we assumed, for now and for simplicity, that this doesn't happen)

Let C and D be the number of concordant and discordant pairs, respectively. The Kendall τ coefficient is defined as:

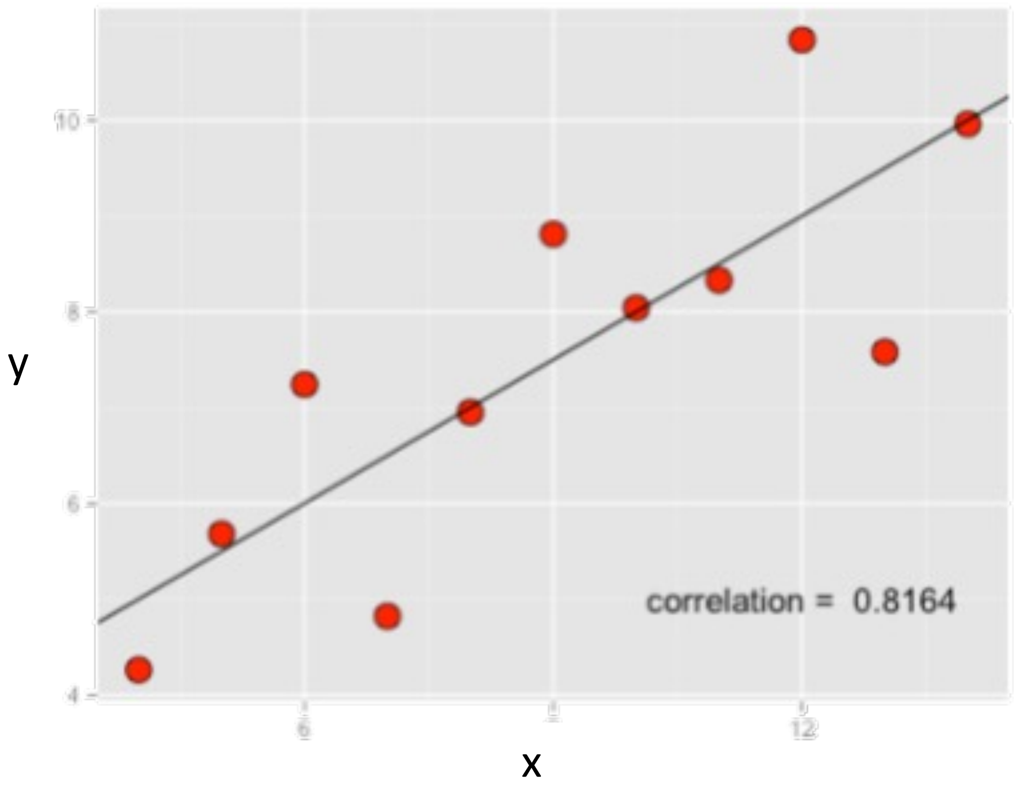
$$\tau = \frac{C - D}{\binom{n}{2}}$$



Kendall correlation coefficient

Let C and D be the number of concordant and discordant pairs, respectively.
The Kendall τ coefficient is defined as:

$$\tau = \frac{C - D}{\binom{n}{2}}$$



How to deal with ties?

$$\tau = \frac{C - D}{\sqrt{\left(\binom{n}{2} - t_x\right)\left(\binom{n}{2} - t_y\right)}}$$

Where t_x and t_y are the number of tied pairs in each of the dimensions, respectively.

Kendall's τ - example

Grades of 11 students in 2 exams:

Exam 1	Exam 2
85	85
98	95
90	80
83	75
57	70
63	65
77	73
99	93
80	79
96	88
69	74

Ranks of exam results and calculating C and D

Exam1 (x)	Exam2 (y)	c	d
1	2	9	1
2	1	9	0
3	3	8	0
4	5	6	1
5	4	6	0
6	7	4	1
7	6	4	0
8	9	2	1
9	8	2	0
10	11	0	1
11	10	**C=50**	**D=5**

$\tau = 0.818$

Statistical assessment (Kendall 1938)

Maurice G Kendall
British statistician
1907-1983

If two permutations are uniformly and independently drawn in S_n then

The Random Variable $\tau = \frac{C-D}{\binom{n}{2}}$ has a standard deviation of $S_\tau = \frac{1}{3}\sqrt{\frac{2n+5}{\binom{n}{2}}}$

and, moreover,

$$Z = \frac{\tau}{S_\tau} = \frac{3(C-D)}{\sqrt{0.5 \cdot n(n-1) \cdot (2n+5)}}$$

has an approximately $N(0,1)$ distribution, for sufficiently large n.

Kendall for the exams data

Exam1 (x)	Exam 2 (y)	C	D
1	2	9	1
2	1	9	0
3	3	8	0
4	5	6	1
5	4	6	0
6	7	4	1
7	6	4	0
8	9	2	1
9	8	2	0
10	11	0	1
11	10		
		C=50	D=5

$$\tau = \frac{50-5}{55} = 0.818$$

$$Z(\tau) = \frac{\tau}{S_\tau},$$

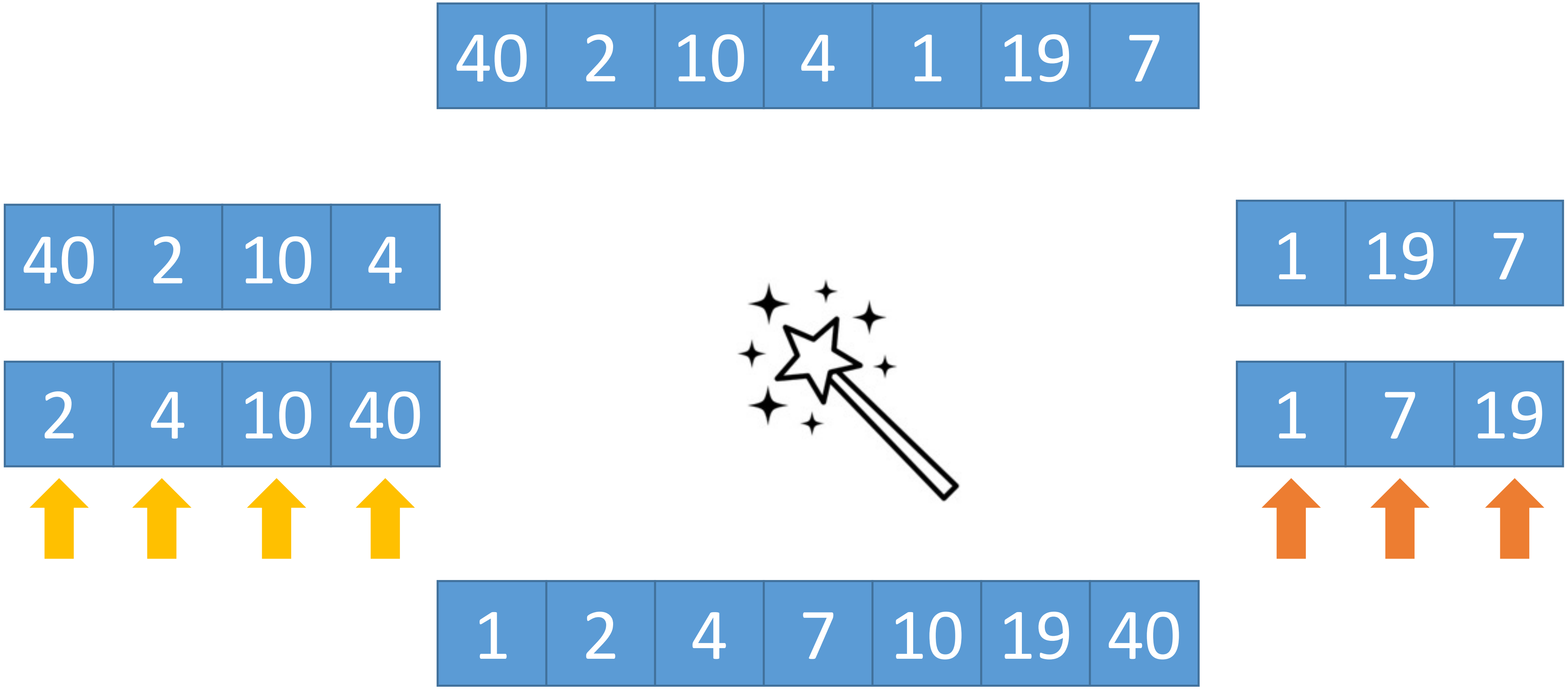
$$S_\tau = \frac{1}{3}\sqrt{\frac{2n+5}{\binom{n}{2}}} = \frac{1}{3}\sqrt{\frac{27}{55}} = 0.23$$

$$Z(\tau) = \frac{0.818}{0.23} \approx 3.5$$

$$\begin{aligned} p-value &= P(Z \geq 3.5) \\ &= 1-\Phi(3.5) \\ &= 1 - 0.0002 \end{aligned}$$

Grades in the two exams are positively correlated w confidence 1 - 0.0002

Merge Sort



Merge Sort

```
Merge_Sort(A)
    Cut A into L and R
    SL = Merge_Sort(L)
    SR = Merge_Sort(R)
    SA = Merge(SL,SR)
        \\ Linearly scan both and insert smaller first
    Return SA
```

Merge Sort

```
def merge_sort(values):
    if len(values) <= 1:
        return values

    m = len(values) // 2
    l, r = values[m:], values[:m]
    l, r = merge_sort(l), merge_sort(r)

    sorted_array = merge(l, r)
    return sorted_array
```

Merge Sort: the Kendall-Knight Algorithm

D_Merge =
When placing an element from the Right, add the number of yet unplaced elements on the Left

(1)	(2)	(3)	(4)	(5)	(6)	(7)
40	2	10	4	1	19	7

40 2 10 4   1 19 7

40 2   10 4   1 19   7

40   2   10   4   1   19   7

D = 1   2 40   D = 1   4 10   D = 0   1 19   D = 0   7

D = DL + DR + D_Merge

2 4 10 40   D = 1   1 7 19

D = (1+1)+1+1 = 4

D = (4 + 1) + 4 + 2 + 1 = 12   1 2 4 7 10 19 40

Kendall-Knight

```
D_Cnt_Merge( L[1..n], R[1..m])
    D=0
    i=0, j=0
    While i ≤ n and j ≤ m Do
        If R[j] < L[i] Then
            D = D+n-i   //Add the number of elements still left in L
            // append R[j] to the sorted array
            j++
        Else
            // append L[i] to the sorted array
            i++
    Return D
```

Complexity analysis

$$T(n) \ = 2T(n/2) + cn$$

$$T(n) \,\epsilon\, O(n \ \log n\ )$$

Summary

- Kendall τ is a rank correlation approach for testing against the uniform permutation rank null model
- Distribution under the null is well characterized
- Kendall-Knight algorithm
- HW:
  + How do Spearman ρ and Kendall τ compare to each other?
  + How do they both compare to Pearson?