Spearman (Inference)

Under the null hypothesis of no association between X_i and Y_i, i.e. rank(X_i) independent of rank Y_i, the sampling distribution of r_s can be found by using permutations. Suppose X_i's are ordered from smallest to largest. The ranks of the Y_i's are a permutation of $1, 2, \ldots, n$. For each ~~permutatio~~ permutation of the Y_i's there ~~e~~ is a corresponding r_s. We have to compute r_s for $n!$ permutation. This works for n small. For n large, an approximation to r_s is used.

Theorem

Under the null hypothesis that the variables are uncorrelated

$$\frac{r_s\sqrt{n-5}}{\sqrt{1-r_s^2}}$$ is approximately t_{n-2} distributed

Kendall

Similary ~~R_s~~ to Spearman, under the null hypothesis of no association between X and Y, we can find the sampling distribution of τ exactly. For n large, approximations are used.

Principal Components Analysis (PCA)

PCA is a type of data compression, i.e. representing data by a smaller amount of data than we are currently using.

Think of the data as forming a matrix

$$\begin{array}{c c} & \begin{matrix} 1 & 2 & \cdots & r \end{matrix} \text{ variates} \\ \text{observations} \begin{matrix} 1 \\ 2 \\ \vdots \\ n \end{matrix} & \left(\begin{matrix} & & & \\ & & & \\ & & & \\ & & & \end{matrix} \right) \end{array}$$

$n \times r$ matrix

rows correspond to observations,
columns correspond to variates,
i.e. characteristics of the observations.

Each observation is a vector in $\mathbb{R}^r$
(no factors allowed)

Some of the variates could be correlated, in which case the data may be described well by fewer coordinates.

Review of Linear Algebra

The dot product of two vectors $\vec{v} = (v_1, \ldots, v_n)$ and $\vec{w} = (w_1, \ldots, w_n)$ is

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} v_i w_i$$

Two vectors are orthogonal if

$$\vec{v} \cdot \vec{w} = 0.$$

A basis for $\mathbb{R}^n$ is a set of vectors $\{\vec{v}_1, \ldots, \vec{v}_n\}$ such that each $\vec{w} \in \mathbb{R}^n$ can be written uniquely as

$$\vec{w} = \sum_{i=1}^{n} \alpha_i \vec{v}_i \quad \text{for } \alpha_i \in \mathbb{R} \ \forall i$$

A ~~baseis~~ basis is called orthogonal if

$$\vec{v}_i \cdot \vec{v}_j = 0$$

i.e. all pairs of basis vectors are ~~orth~~ orthogonal.

If, in addition, $v_i \cdot v_i = 1 \ \forall i$, then the basis is orthonormal. Each ~~vector~~ basis vector has unit length.

The standard orthonormal basis for $\mathbb{R}^n$ is

$\{e_1, \ldots, e_n\}$

where $e_i = 1$ at position i

$= 0$ at position $j \neq i$

$e_1 \quad (1, 0, 0 \ldots)$

$e_2 \quad (0, 1, 0 \ldots)$

$e_3 \quad (0, 0, 1, \ldots)$

$\vdots$