Numerical data description

Non-robust vs. robust measures

(E.g. measurement error)

Location: "where"

Sample mean	not robust
Median	robust
Mode	robust

[Quantiles, quartiles]

Scale: "spread"

Sample variance	not robust
Sample standard deviation	not robust
Median absolute deviation	robust
Interquartile range	robust

Part of Exploratory Data Analysis:

Understanding of the data structure

Shape measures

- Sample skewness:

$$\text{skew}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\text{std}(x)}\right)^3$$

not robust

standardization

Measure asymmetry of data

Long right tail: skew > 0

Long left tail: skew < 0

Symmetric around mean $\Rightarrow$ skew = 0

- Galton's measure of skewness: robust

$$\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

Normalised difference between ranges of centre quartes

- Sample kurtosis:

$$\text{kurt}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\text{std}(x)}\right)^4$$

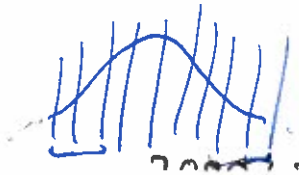
not robust

Heaviness of tails: ignores data close to mean

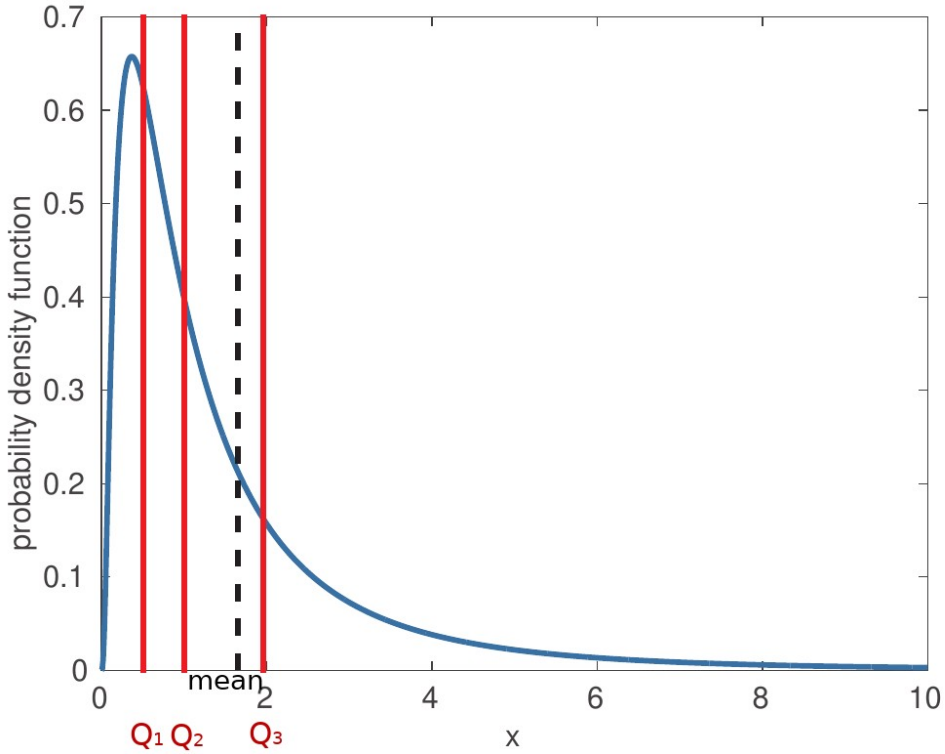
- Robust kurtosis:

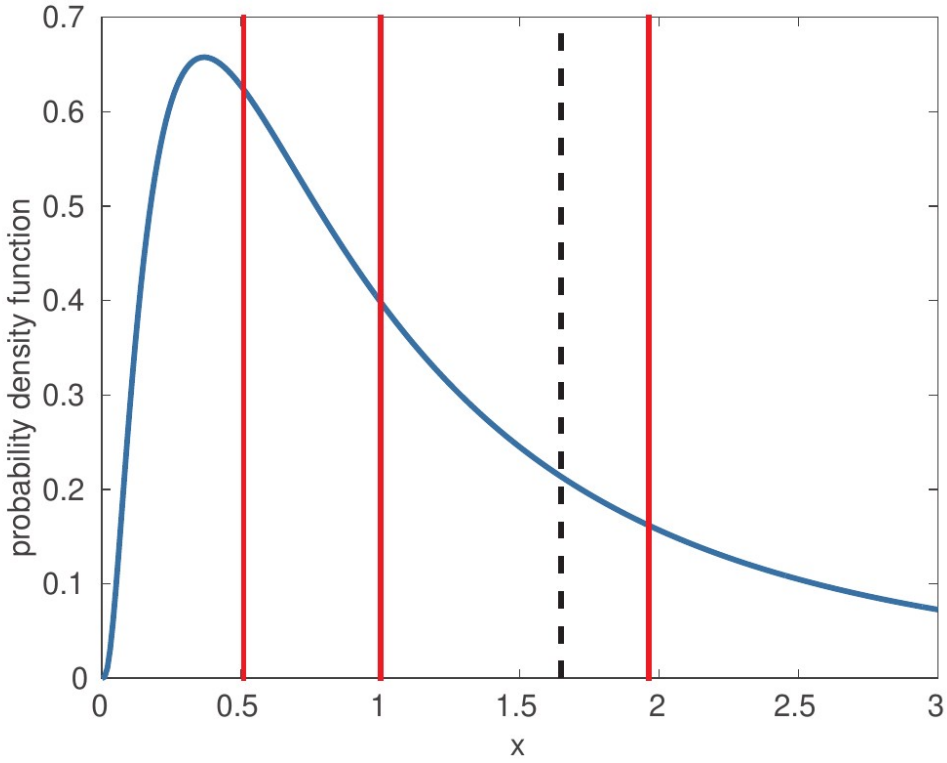
$$\frac{(q_{7/8} - q_{5/8}) + (q_{3/8} - q_{1/8})}{q_{3/4} - q_{1/4}}$$

robust

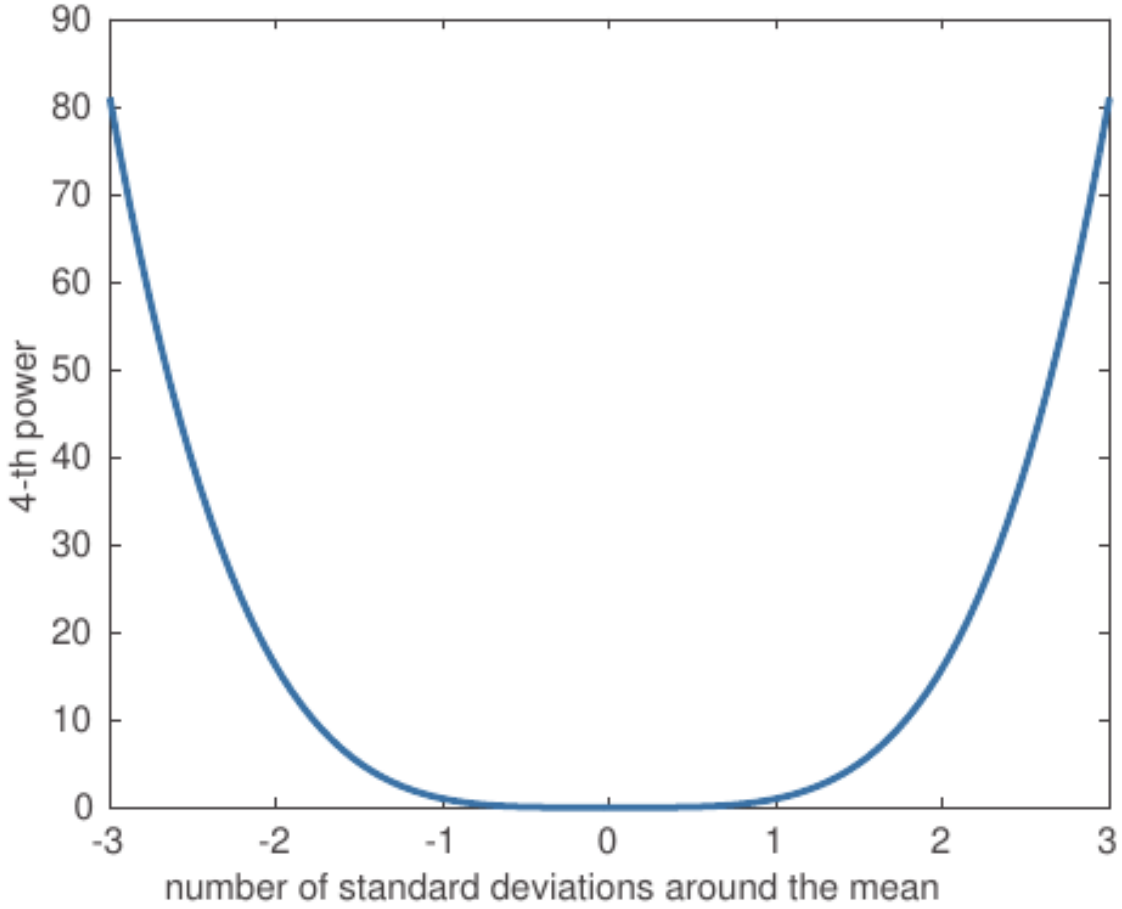


Positive Skewness





Fourth Power



Dependence measures

Observations: $x = (x_1, \dots, x_n)$ $y = (y_1, \dots, y_n$

- Sample covariance: not robust

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Estimate of covariance

$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Strength of linear association

It holds:

$$\text{Cov}[X, X] = \text{Var}[X]$$
$$\text{Cov}[X, Y] = \text{Cov}[Y, X]$$
$$\text{Cov}[aX + b, Y] = a\,\text{Cov}[X, Y]$$

$\rightsquigarrow$ scale dependent

- Pearson's correlation coefficient: not robust

$$\rho(x,y) = \frac{cov(x,y)}{std(x)\,std(y)}$$

Normalises covariance

Measures linear relationship:

Let $y = ax + b$, $a \neq 0 \rightsquigarrow \bar{y} = a\bar{x} + b$

$$std(y) = \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2} = \sqrt{\frac{1}{n}\sum (ax_i + b - a\bar{x} - b)^2}$$

$$= \sqrt{a^2}\, std(x) = |a|\, std(x)$$

$$cov(x,y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$

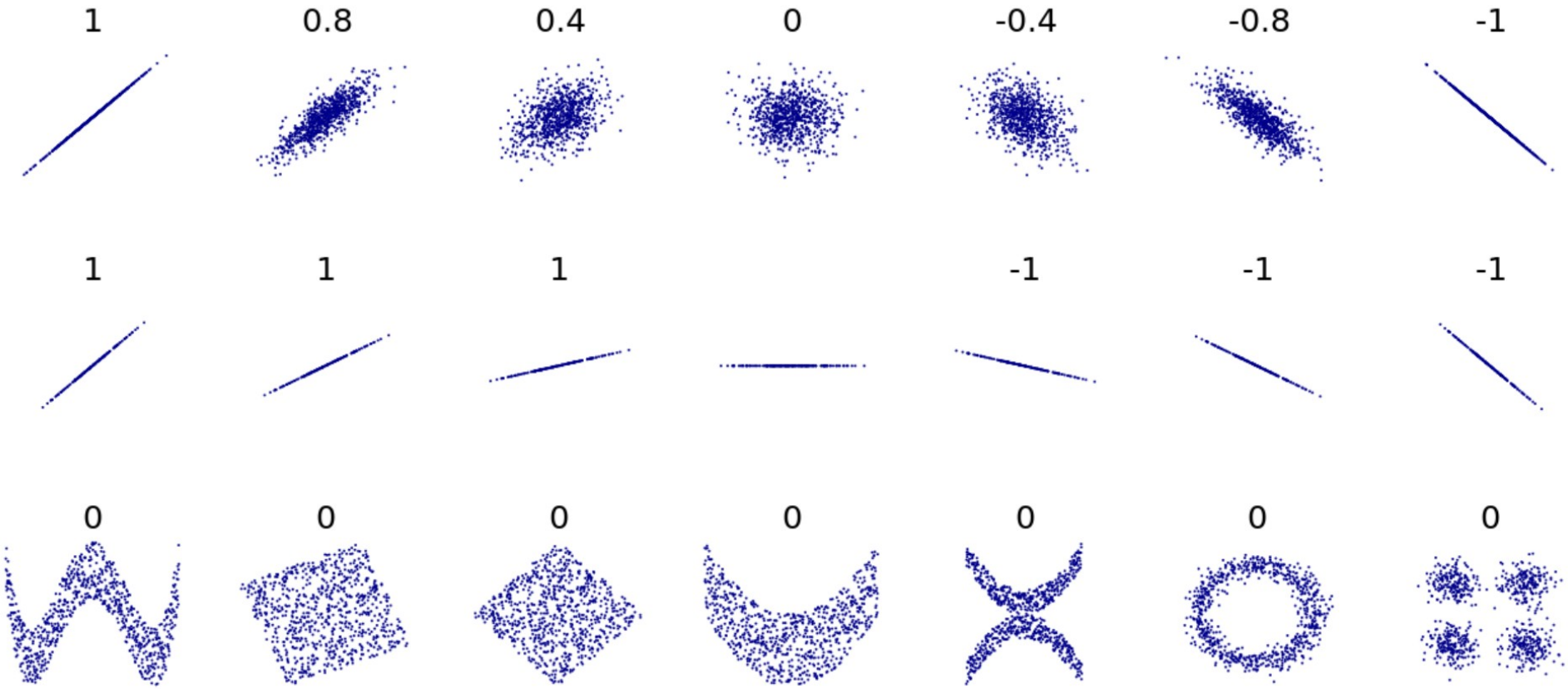
$$= \frac{1}{n}\sum (x_i - \bar{x})(ax_i + b - a\bar{x} - b)$$

$$= a\, var(x)$$

$$\rho(x,y) = \frac{cov(x,y)}{std(x)\,std(y)} = \frac{a\, var(x)}{std(x)|a|\, std(x)}$$

$$= \frac{a}{|a|} = \begin{cases} 1 & ; a > 0 \\ -1 & ; a < 0 \end{cases}$$

Uncorrelated and Dependent



Source: https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

- Kendall's τ: robust

$$\tau(x, y) = \frac{n_c(x, y) - n_d(x, y)}{n(n-1)/2}$$

where:

n = # observations
("Number of")

n_c = # concordant pairs
$(x_i, y_i), (x_j, y_j), i \neq j$:
$x_i > x_j$ and $y_i > y_j$
or $x_i < x_j$ and $y_i < y_j$
"element order consistent"

n_d = # discordant pairs
$(x_i, y_i), (x_j, y_j), i \neq j$
$x_i > x_j$ and $y_i < y_j$
or $x_i < x_j$ and $y_i > y_j$
"element order inconsistent"

Rank correlation measure:
order of elements important