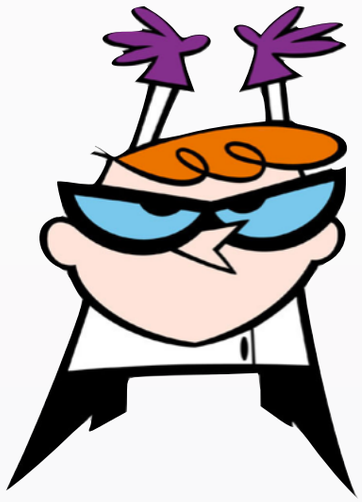
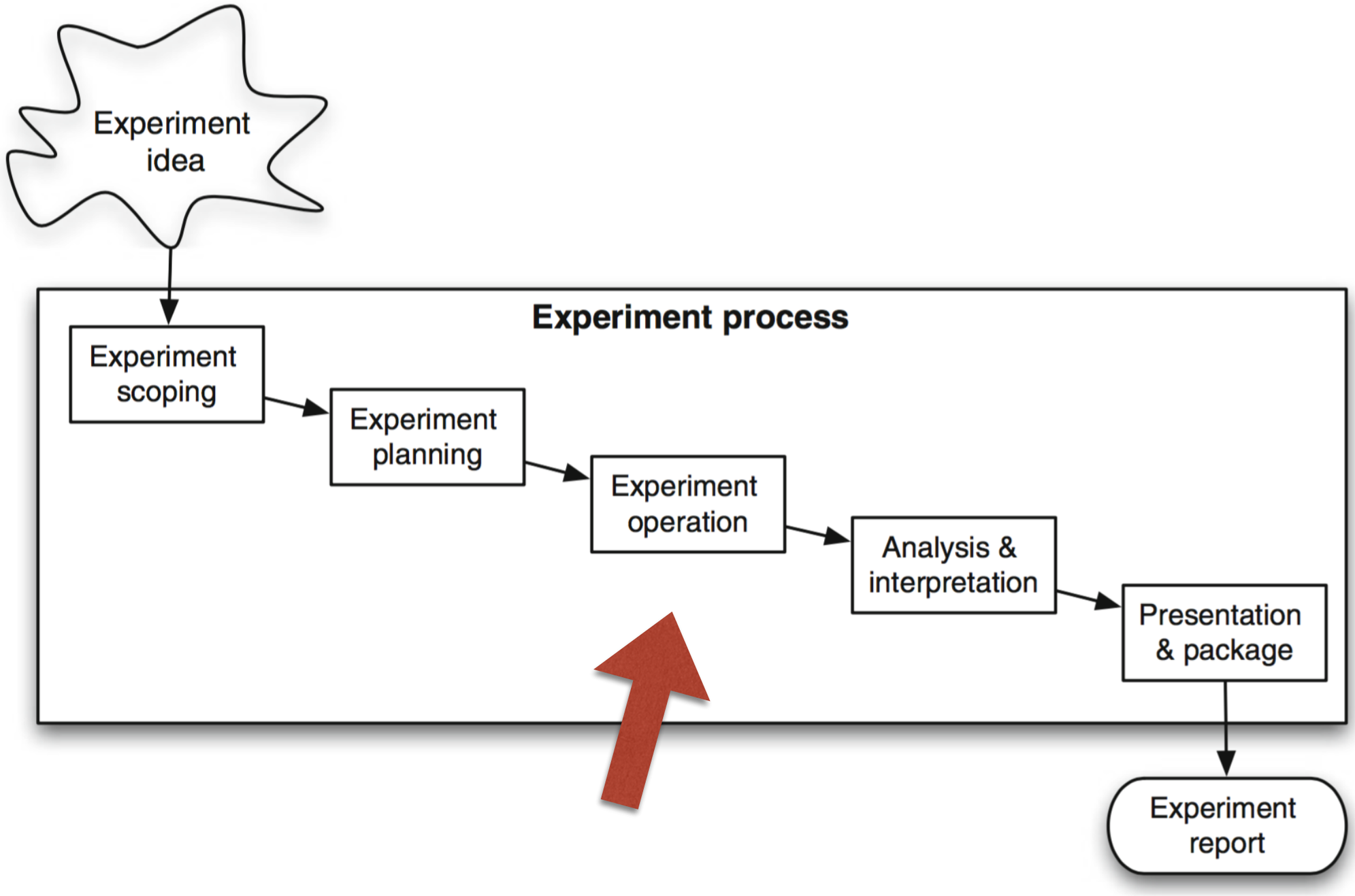
2. experimentation

things we learned so far

- design of an experiment
  - variables, subjects, objects, factors, treatments, tests
- statement of purpose
- hypothesis formulation (null and alternative)
- 3 experimental design principles
- standard design types
- validation threats

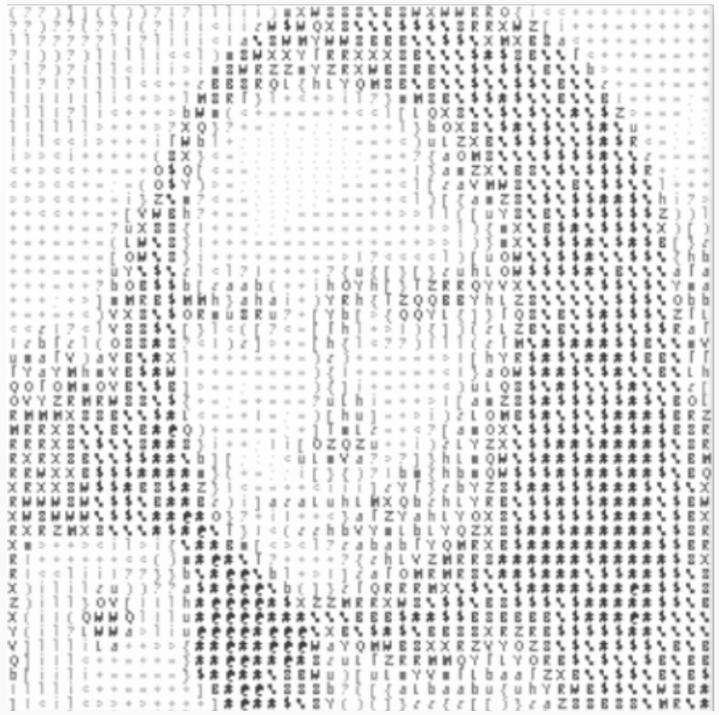




Source: Experimentation in Software Engineering, Springer, 2012

quantitative interpretation

1. descriptive statistics
   - get a feeling of how the data are distributed
2. reducing the data set
   - systematic errors vs. outliers
3. hypothesis testing
   - aka “can we reject H_0?”



scale types

- **Nominal**: maps the attribute to a name or symbol
  - examples: eye colour, religion, sex
- **Ordinal**: ranks entities after an ordering criterion
  - examples: grades, level of agreement, rank
- **Interval**: possesses equal intervals
  - examples: temperature in Celsius and Fahrenheit
- **Ratio**: has an absolute zero, the ratio is meaningful
  - examples: length, temperature in Kelvin, income

Source: http://psychology.ucdavis.edu/sommerb/sommerdemo/scaling/levels.htm

time of day: example

- **Nominal**

  - categories; no additional information

- **Ordinal**

  - indicates order of occurrence; spacing is uneven

- **Interval**

  - equal intervals; analog (12-hr.) clock

- **Ratio**

  - 24-hr. time has an absolute 0 (midnight)

Source: http://psychology.ucdavis.edu/sommerb/sommerdemo/scaling/levels.htm

measures of central tendency

- the “middle” of the data set
- estimation of the expectation of the stochastic variable from which data points are sampled.

tendency: mean

- given $X = \{x_1, x_2, \ldots x_n\}$
  - $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- Example. X = {18, 3, 6, 4, 8, 9}
  - $\bar{x}$ = 48/6 = 8

tendency: median

- given a **sorted** $X = \{x_1, x_2, \ldots x_n\}$

- $$\mu_{i/2} = \begin{cases} x_{i/2} & n - odd\ number \\ \frac{x_{i/2-1} + x_{i/2}}{2} & n - even\ number \end{cases}$$

- Example. X = {18, 3, 6, 4, 8, 9}
  - X_{sorted} = {3, 4, 6, 8, 9, 18}
  - median = (6+8)/2 = 7

tendency: percentile

- x_p denotes the percentile where p% of the **sorted** samples lies below this value.
- Example. X = {18, 100, 3, 1, 11, 45, 6, 4, 8, 9}
  - X_{sorted} = {1, 3, 4, 6, 8, 9, 11, 18, 45, 100}
  - $X_{25\%} = 6$; $X_{50\%} = 9$; $X_{75\%} = 45$; $X_{99\%} = 100$

tendency: mode

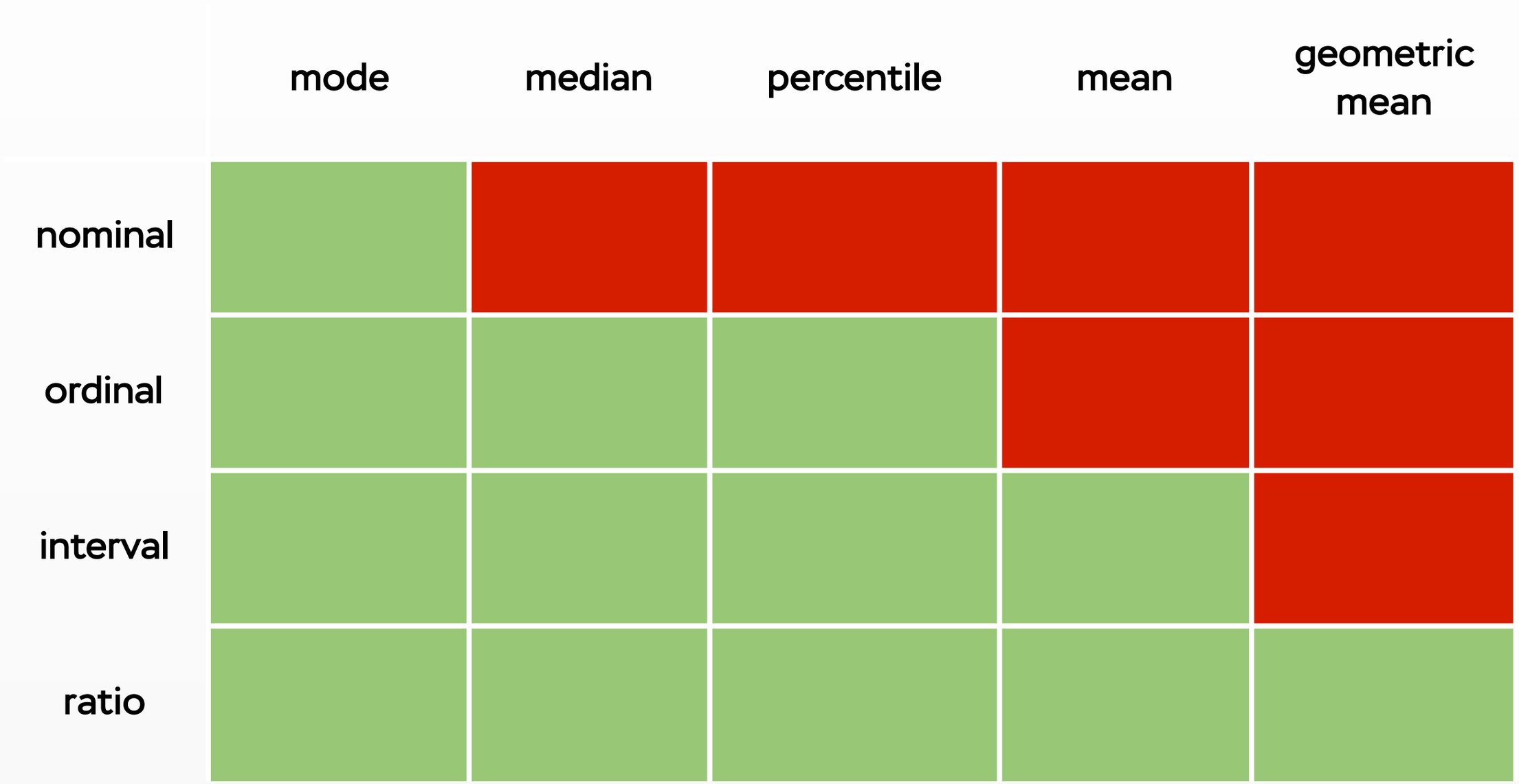
- the most commonly occurring sample
- Example. X = {18, 100, 3, 1, 1, 45, 6, 4, 8, 9}
  - mode = 1
- Example. X = {18, 100, 3, 1, 1, 4, 6, 4, 8, 9}
  - mode = ?

tendency: geometric mean

- given $X = \{x_1, x_2, \ldots x_n\}$
- $\sqrt[n]{\prod_{i=1}^{n} x_i}$
- Example. X = {18, 3, 6, 4, 8, 9}
  - $\bar{x}$ = n_{th}root(93312, 6) = 6.734772289856237

Note: All the samples must be non-negative

tendency measures vs. scales



measures of dispersion

- level of variation from the central tendency
- how spread or concentrated the data is

dispersion: variance

- given $X = \{x_1, x_2, \ldots x_n\}$
- $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
- Example. X = {18, 3, 6, 4, 8, 9}
  - $\bar{x}$ = 48/6 = 8
  - $s^2 = [10^2 + 5^2 + 2^2 + 4^2 + 0^2 + 1^2]/5 = 24.333333\ldots$

dispersion: standard deviation

- given $X = \{x_1, x_2, \ldots x_n\}$
- $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
- Example. X = {18, 3, 6, 4, 8, 9}
  - $\bar{x}$ = 48/6 = 8
  - s = sqrt{[10^2 + 5^2 + 2^2 + 4^2 + 0^2 + 1^2]/5} = 4.9328…

dispersion: range

- given X = {x_1, x_2, … x_n}
  - range = max(X) - min(X)

- Example. X = {18, 3, 6, 4, 8, 9}
  - range = 18 - 3 = 15

dispersion: variation interval

- given $X = \{x_1, x_2, \ldots x_n\}$
  - var. interval = (min(X), max(X))

- Example. $X = \{18, 3, 6, 4, 8, 9\}$
  - var. interval = (3, 18)

Note: minimum and maximum are included

dispersion: coeff. of variation

- given X = {x_1, x_2, … x_n}
  - coeff. of variation = $100\frac{s}{\bar{x}}$
- Example. X = {18, 3, 6, 4, 8, 9}
  - s = 4.9328
  - $\bar{x}$ = 8
  - coeff. of variation = 100*4.9328/8 = 61.66%

dispersion: frequency

- given $X = \{x_1, x_2, \ldots x_n\}$
- $f(x_i)$ = frequency of x_i
- $f(x_i)/n$ = relative freq.

value	frequency	relative frequency
1	3	0.23
2	2	0.15
3	1	0.08
4	3	0.23
5	1	0.08
6	2	0.15
7	1	0.08

meas. of dispersion vs. scales

	frequency	variation interval	variance	standard deviation	range	coefficient of variation
nominal	✓	✗	✗	✗	✗	✗
ordinal	✓	✓	✗	✗	✗	✗
interval	✓	✓	✓	✓	✓	✗
ratio	✓	✓	✓	✓	✓	✓

measures of dependency

- indicate how “dependent” X and Y are
  - X and Y being 2 stochastic variables
- samples in pairs (x_i, y_i)

dependency: comm. formulae

- $$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$
- $$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
- $$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \left(\sum_{i=1}^{n} (x_i - y_i) \right) - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right)$$

dependency: linear regression

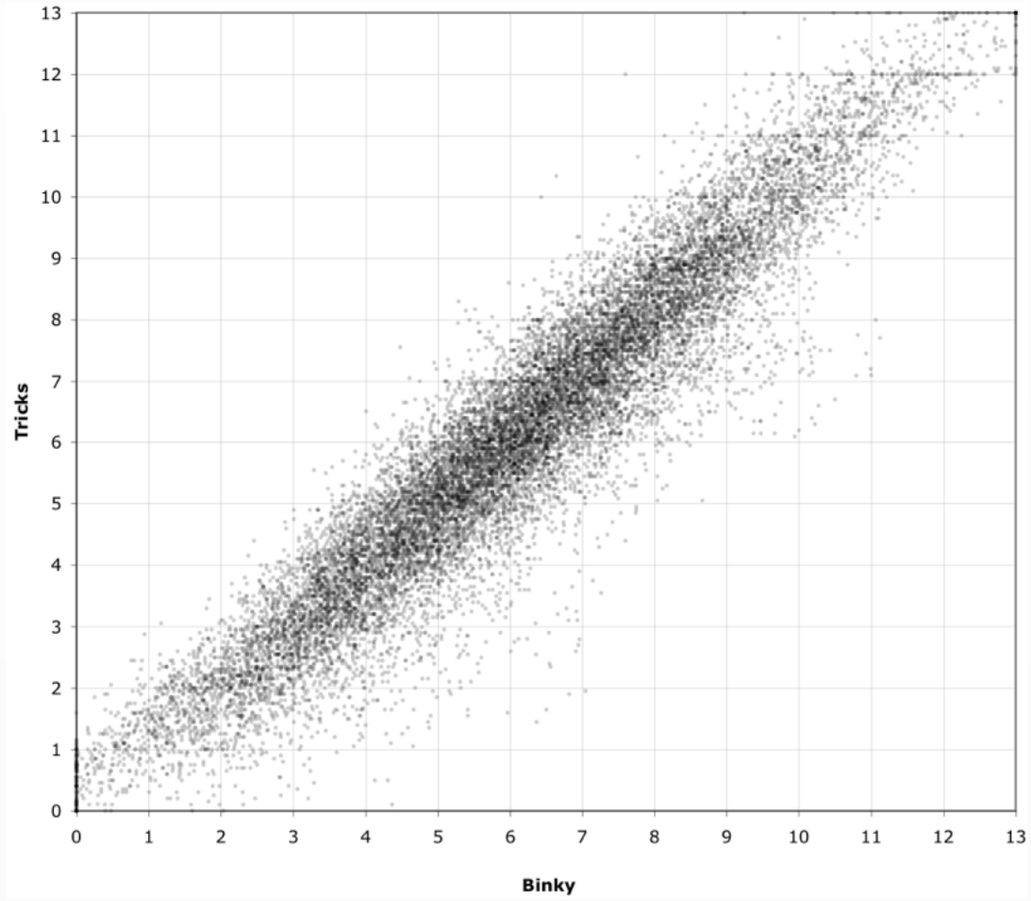
- given $X = \{x_1, x_2, \ldots x_n\}$

  and $Y = \{y_1, y_2, \ldots y_n\}$

- if we suspect that y = f(x), where f(x) is linear

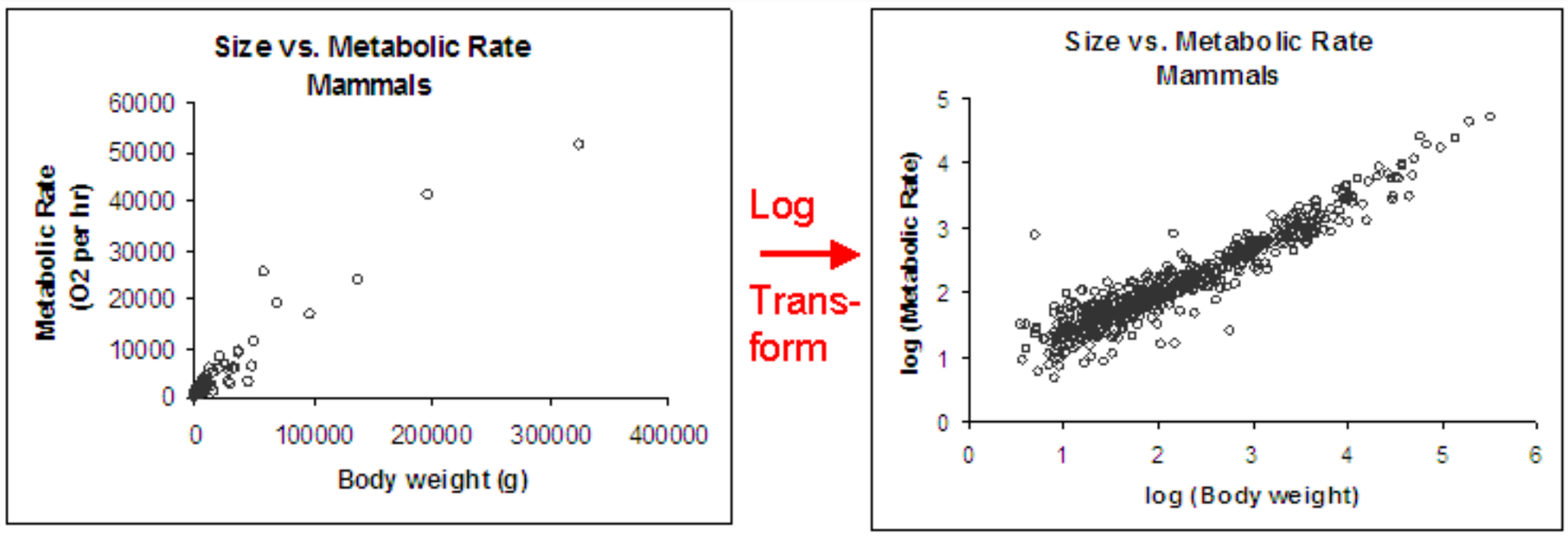
- $y=\alpha+\beta x$,

  where $\alpha = \overline{y} - \beta\overline{x}$ and $\beta = \frac{S_{xy}}{S_{xx}}$



dependency: transformation

- if we suspect that f(x) is not linear, we can apply a transformation to the data and than use linear regression.
- Exponential relation? > logarithmic transformation



dependency: covariance

- given $X = \{x_1, x_2, \ldots x_n\}$ and $Y = \{y_1, y_2, \ldots y_n\}$
- $$c_{xy} = \frac{S_{xy}}{n-1} \quad \in (-\infty, +\infty)$$

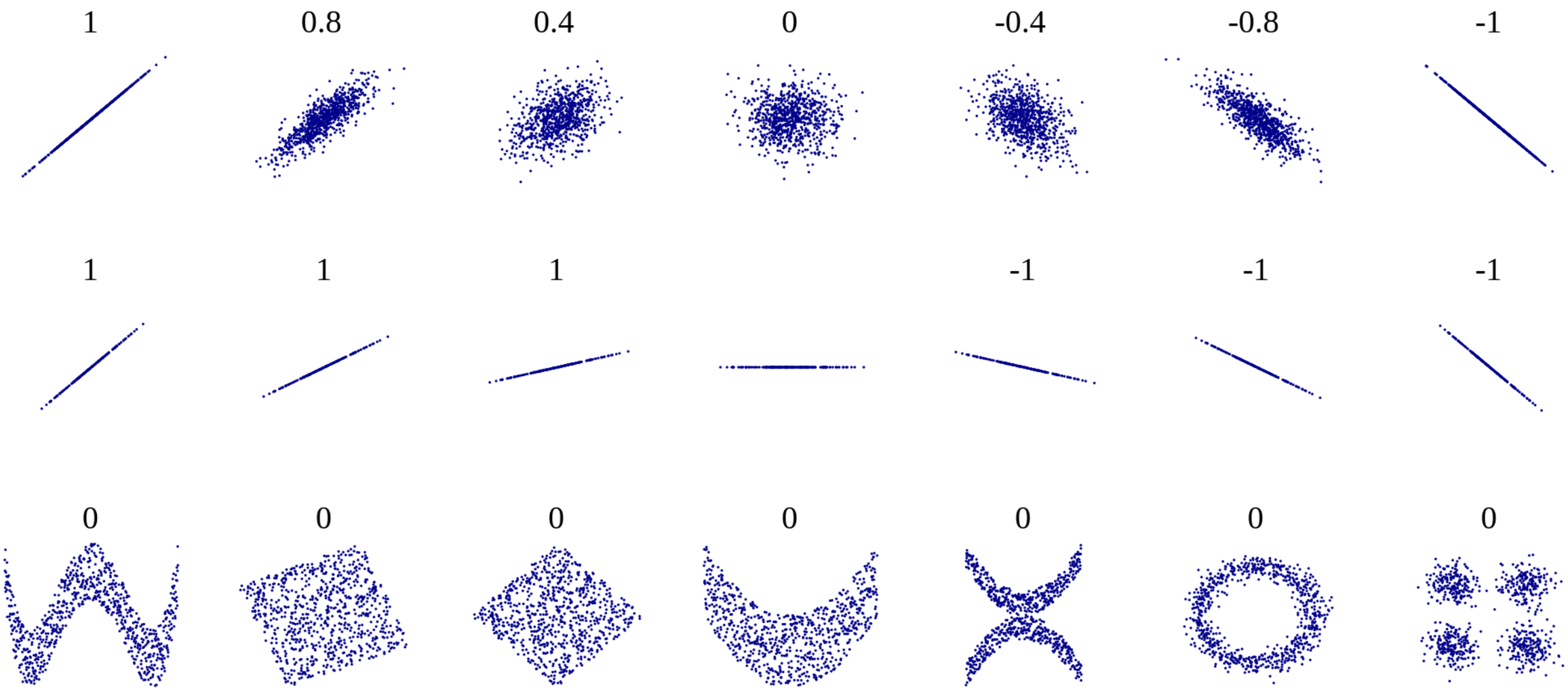
- Example. $X = \{2.1, 2.5, 4.0, 3.6\}$, $Y = \{8, 12, 14, 10\}$
  - $c_{xy} = 1.53$
  - positively related

depend.: Pearson corr. coeff.

- given $X = \{x_1, x_2, \dots x_n\}$ and $Y = \{y_1, y_2, \dots y_n\}$
- $$r = \frac{S_{xy}}{S_x S_y} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \in [-1, 1]$$
- Example. X = {2.1, 2.5, 4.0, 3.6}, Y = {8, 12, 14, 10}
  - $c_{xy} = 0.66$
  - positively related

Note: It spots linear correlations only

Pearson's known limitations



depend.: Spearman corr. coeff.

- given X = {x_1, x_2, … x_n} and Y = {y_1, y_2, … y_n}

  in ordinal scale or far from being normally distributed

  - $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$ $\in$ [-1, 1]

- Example.

  IQs = {106, 86, 100, 101, 99, 103, 97, 113},

  hs of TV per week = {7, 0, 27, 50, 28, 29, 20, 12}

  - ρ = 0.11905

Note: It spots linear correlations only

meas. of dependency vs. scales