データの散らばりの数値化と分析

変量 (X)

計測対象の特性(プロパティ)
- 数学の試験 …… 点数
- 身体測定 …… 身長，体重
- さいころ …… 目の値
- 天気 …… 気温，降水量

データ (data)

変量の測定値の集まり

$x_1, x_2, x_3, \cdots\cdots, x_n$

（データの大きさ n）

変量 X のデータ

$x : x_1, x_2, x_3, \cdots\cdots, x_n$

の散らばりの数値化ステップ

Step1

データの平均値 $\overline{x}$ を求める

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Step2

平均からの散らばり(偏差)

$x_1 - \overline{x},\ x_2 - \overline{x},\ x_3 - \overline{x},\ \cdots,\ x_n - \overline{x}$

偏差の和は 0 だから，和で散らばりは表せない

Step3

散らばりを偏差の大きさの和で数値化

○絶対値で散らばりを表す

$$|x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}|$$

○２乗して散らばりを表す

$$(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2$$

２つの変量のデータの個数が違うと，散らばりの比較はできない。

Step4

データの個数で割り，散らばりを揃える(平均化)

○平均偏差(*Mean Deviation*)

$$MD = \frac{|x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}|}{n}$$

○分散(*Variance*)

$$s_x^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}$$

分散の単位はデータの単位の２乗になってしまう
例えば，長さ(cm)の分散は面積(cm²)

Step5

データと分散の散らばりの単位を揃える

○標準偏差(*Standard Deviation*)

$$s_x = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}}$$

偏差を計算しないで標準偏差(分散)は求められる

○標準偏差の簡便式

$$s_x = \sqrt{\overline{x^2} - (\overline{x})^2}$$

○偏差値(*Standard Score*)

$$ss = \frac{x - \overline{x}}{s_x} \times 10 + 50$$

目盛の変換

$s \Rightarrow 1 \Rightarrow 10$

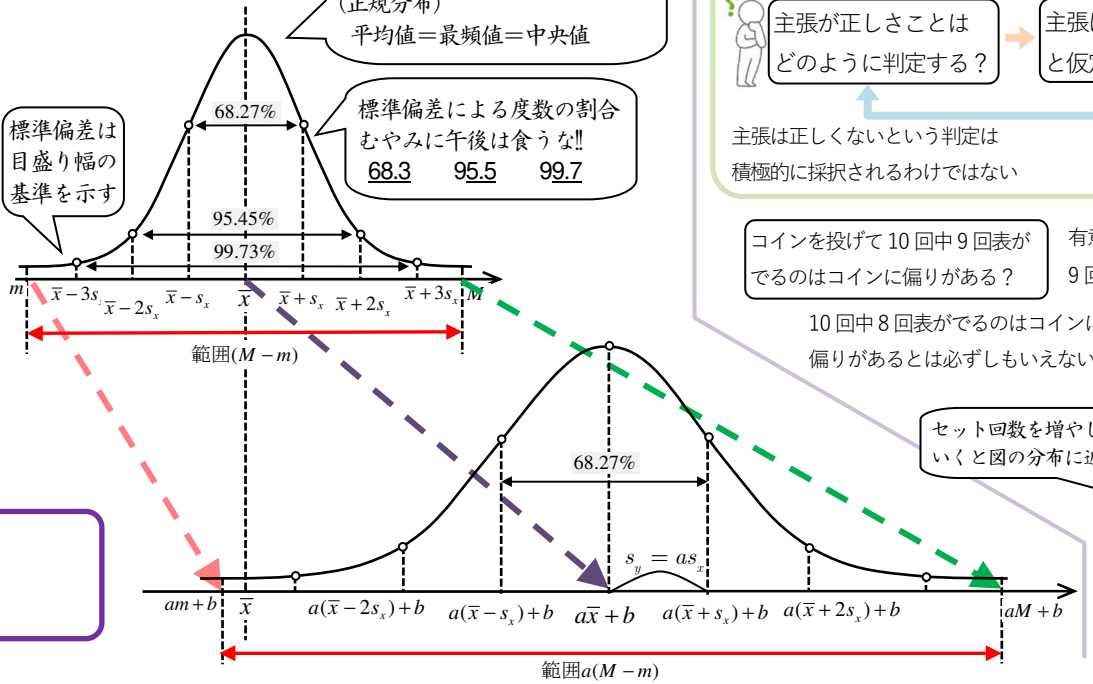
平均の変換

$\overline{x} \Rightarrow x - \overline{x}(0) \Rightarrow 50$

変量のデータの個数が多くなるとデータの分布は釣鐘状になる(正規分布)
平均値＝最頻値＝中央値

標準偏差による度数の割合
むやみに午後は食うな!!
68.3　95.5　99.7

標準偏差は目盛り幅の基準を示す



仮説検定

有意水準(危険率)付きの背理法

対立仮説 H_1　帰無仮説 H_0

主張が正しさことはどのように判定する？ → 主張は正しくないと仮定する → 公平な実験を繰り返し統計量(度数)を調べる → 有意水準による判定

帰無仮説 H_0 は棄却される(無に帰する) → 棄却域

帰無仮説 H_0 は採択される → 受容域

主張は正しくないという判定は積極的に採択されるわけではない

コインを投げて 10 回中 9 回表がでるのはコインに偏りがある？

有意水準を $\alpha = 5\%\ (= 0.05)$ とする．

9 回以上の相対度数は $\dfrac{10+1}{1000} = 0.011 < \alpha$

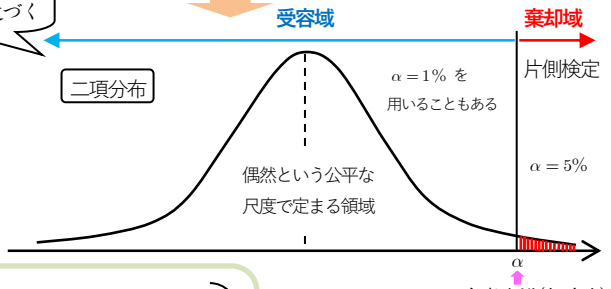
10 回中 9 回表がでるのはコインに偏りがある

10 回中 8 回表がでるのはコインに偏りがあるとは必ずしもいえない

公正なコインを 10 回投げて表のでた回数を記録する実験を 1000 セット

表の回数	0	1	2	3	4	5	6	7	8	9	10	計
度　数	1	9	42	115	204	246	211	117	44	10	1	1000

セット回数を増やしていくと図の分布に近づく



変量の変換

変量 X のデータ x を a 倍し b を加えるとデータの散らばりはどう変化するだろう

○a 倍する$(a>0)$
平均と標準偏差は a 倍になる

○b を加える
平均は b 増える。散らばりは度数分布の階級がズレるだけなので変わらない。

$x \to y = ax + b$ とすると

平均　$\overline{y} = a\overline{x} + b$

標準偏差　$s_y = |a| s_x$

$\Rightarrow s_y^2 = a^2 s_x^2$

散らばりは正の値

仮平均 x_0，$u = \dfrac{x - x_0}{c}$ とする

$\overline{u} = \dfrac{\overline{x} - x_0}{c}$，$s_u^2 = \dfrac{1}{c^2} s_x^2$

Fuminori Nakamura