データ解析 第7回
母集団と標本

北九州市立大学経済学部

齋藤 朗宏

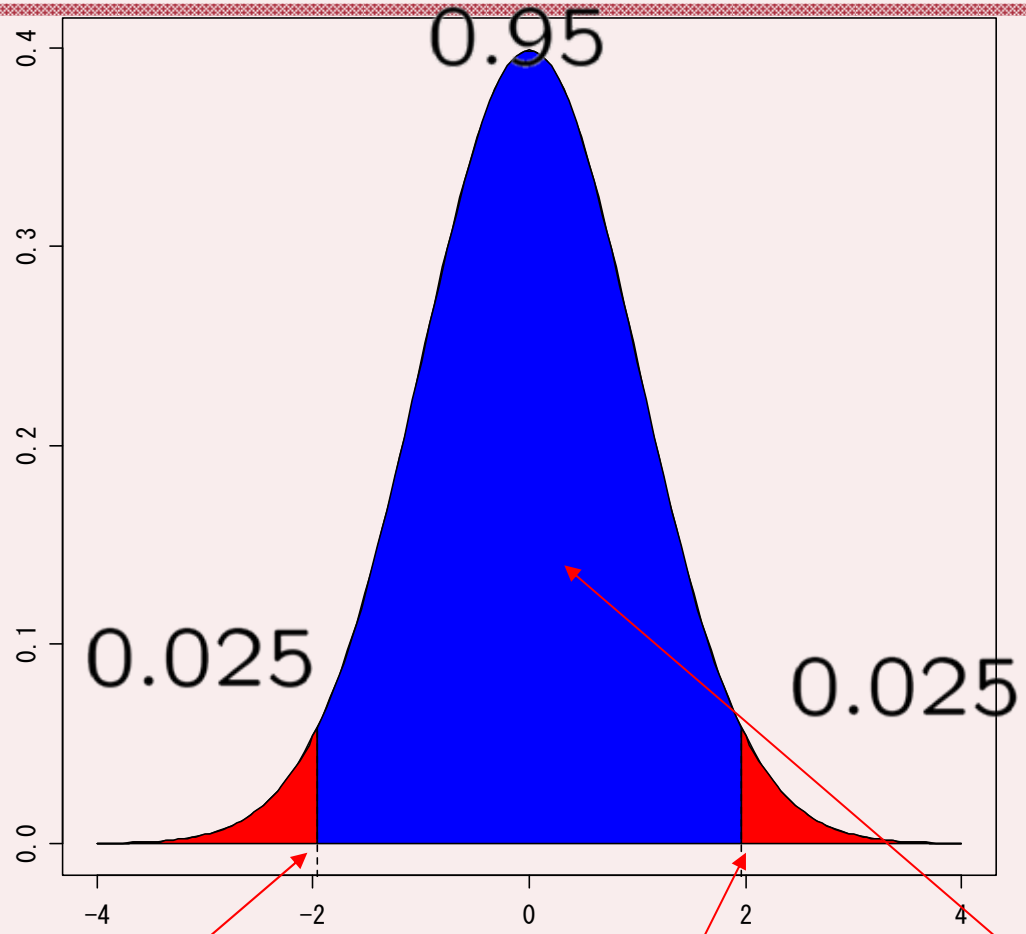
今日の話題

- 正規分布と標準正規分布
- 母集団と標本
- 標本統計量

- 実習

正規分布と標準正規分布

標準正規分布における確率



- 標準正規分布において，以下の関係が成立する．

$$P(-1.96 \leq z_i \leq 1.96) = 0.95$$

正規分布における確率

- 標準化の考え方から，以下の関係が成立する.

$$P(-1.96 \leq z_i \leq 1.96) = 0.95$$

$$z_i = \frac{x_i - \mu_x}{\sigma_x}$$

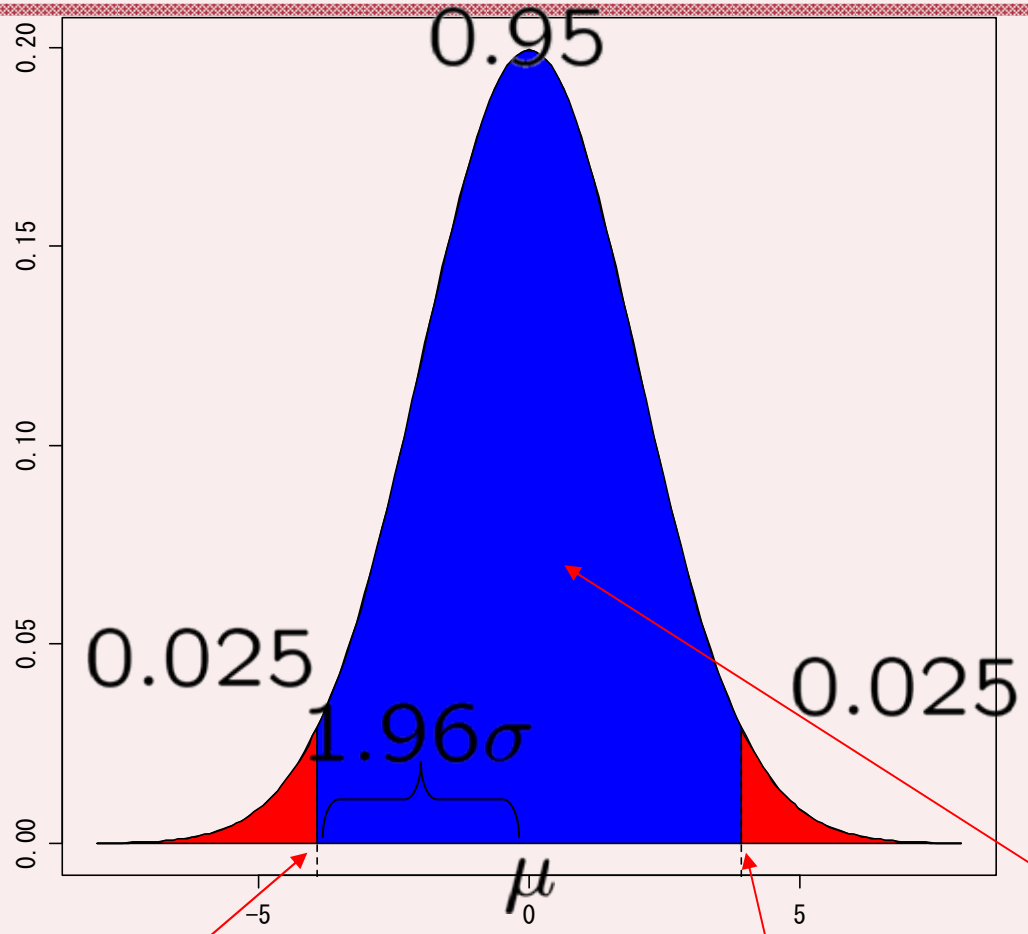
$$P(-1.96 \leq \frac{x_i - \mu_x}{\sigma_x} \leq 1.96) = 0.95$$

$$P(-1.96\sigma_x \leq x_i - \mu_x \leq 1.96\sigma_x) = 0.95$$

$$P(\mu_x - 1.96\sigma_x \leq x_i \leq \mu_x + 1.96\sigma_x) = 0.95$$

即ち，x_iが$\mu_x \pm 1.96\sigma_x$の範囲に入る確率は0.95.

正規分布における確率

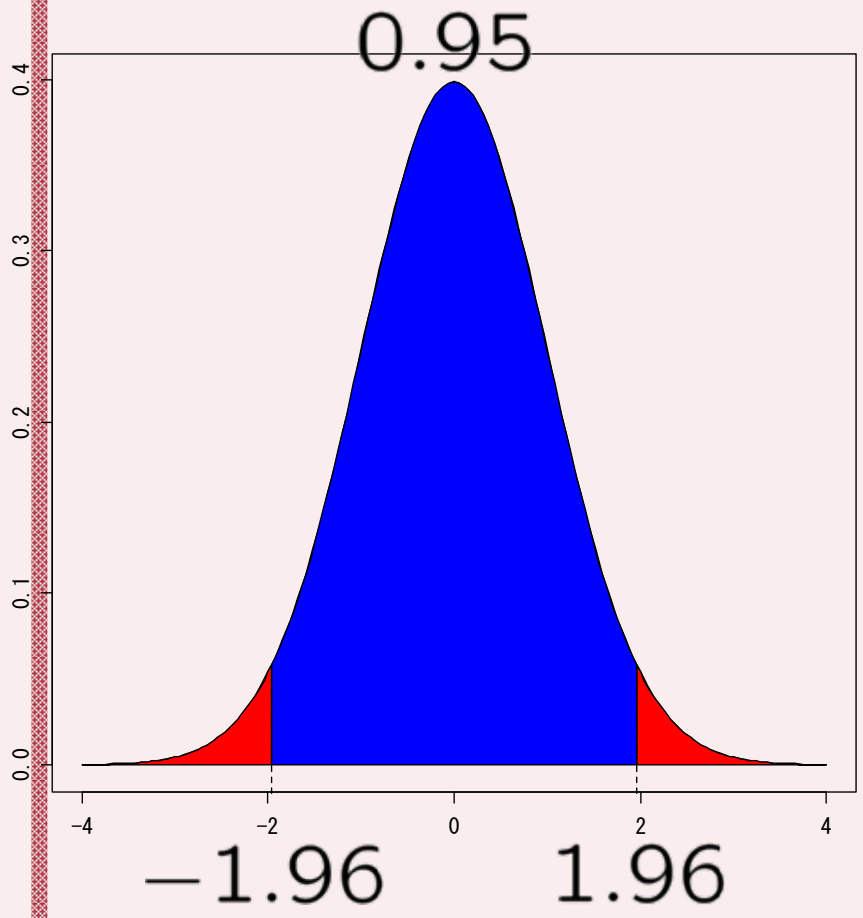


- 正規分布においては，以下の関係が成立する．

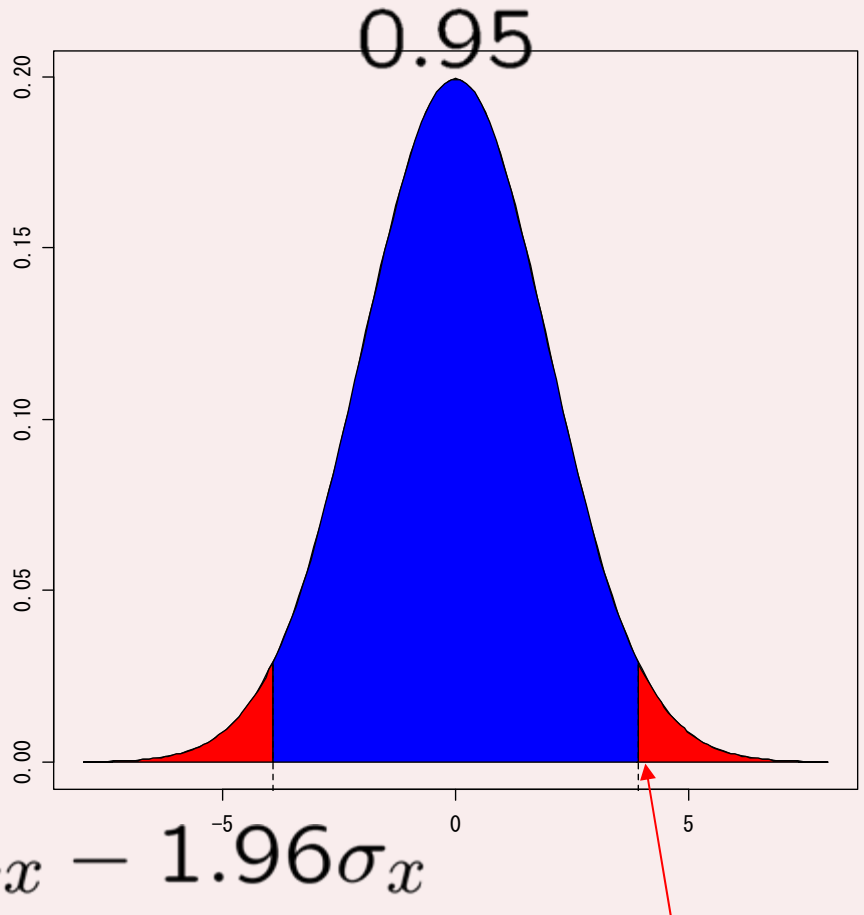
$$P(\mu_x - 1.96\sigma_x \leq x_i \leq \mu_x + 1.96\sigma_x) = 0.95$$

標準正規分布と正規分布

標準正規分布



正規分布



標準正規分布と正規分布の関係

- 正規分布に従う確率変数x と標準正規分布に従う確率変数 z があったとする．このとき，以下の関係が成立する．

$$P\left(\frac{x_i - \mu_x}{\sigma_x} \leq a\right) = P\left(z_j \leq a\right)$$

- 標準化を行えば，「全体の95%は±1.96の範囲に収まる」という形で，分布に関する評価が簡単になる．

標本統計量

標本抽出

- 調査対象となる集団のことを母集団（Universe, Population）と呼ぶ（ex. 日本人，大学生）．
- 母集団のすべての個体に対して行う調査を全数調査と呼ぶ．
- 母集団における統計量を母数と呼ぶ（母平均）．
- 母集団全体に対して調査を行うのは難しいので，母集団の中から一部の個体（標本）を選んで（抽出・サンプリング）調査（標本調査）・測定し，母数の推定値とする．サンプリングは，原則ランダムであるべき（無作為抽出）．
- 標本から計算された母数の推定値を標本平均などと呼ぶ．

母平均と標本平均

- 「日本人の身長」が知りたいとする.
  - 母集団は日本人全体となる．日本人全員の身長を測定し，その平均を求めると，その結果は母平均 μ_x となる.
  - 日本人の中からランダムに何人か選び，その身長の平均を求めると，得られる値は標本平均 $\overline{x}$ となる.

- 「北九大生のTOEICの得点」が知りたいとする.
  - 母集団は北九大生全体．北九大生全員の得点を調べ，その平均を求めると，その結果は母平均 μ_y .
  - 北九大生の中からランダムに何人か選び，その得点の平均を求めると，得られる値は標本平均 $\overline{y}$ となる.

母平均・標本平均

- 母集団の総数をN，標本数をnとする．このとき，確率変数xの母平均μ_xは以下の通り．

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- 標本平均$\bar{x}$は以下の通り．

$$\hat{\mu}_x = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

母分散・標本の分散

- 母分散 σ_x^2 は以下の通り.

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

- 標本の分散$\hat{\sigma}_x^2$は以下の通り.

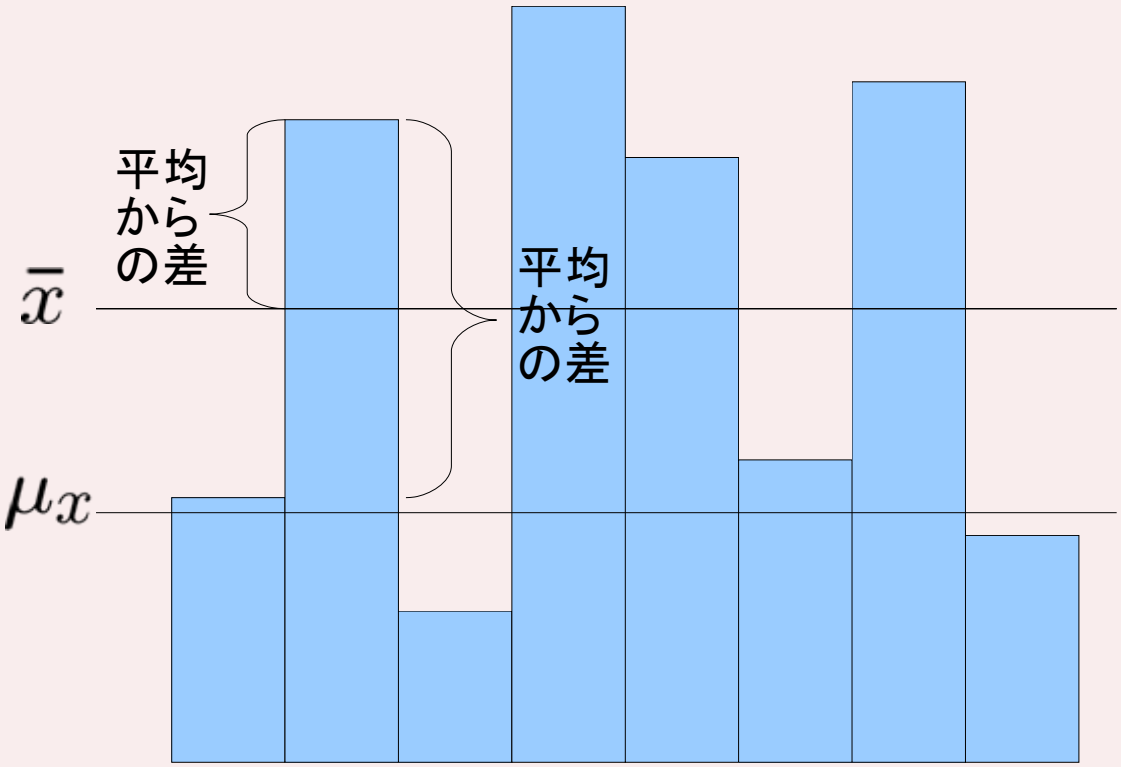
$$\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2$$

手元には標本しかない．よって母平均は未知

$$\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

これにも問題がある

何故不偏分散か？



- $\bar{x}$ と μ_x とは多くの場合一致しない，μ_x の真の値はどこだかわからないが，それが少しでも$\bar{x}$とずれていた場合，標本分散では過小推定．

不偏分散

- 母分散 σ_x^2 は以下の通り.

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

- 不偏分散 s_x^2 は，母集団数 N が十分大きい時以下の通り.

$$\hat{\sigma}_x^2 = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

従って，

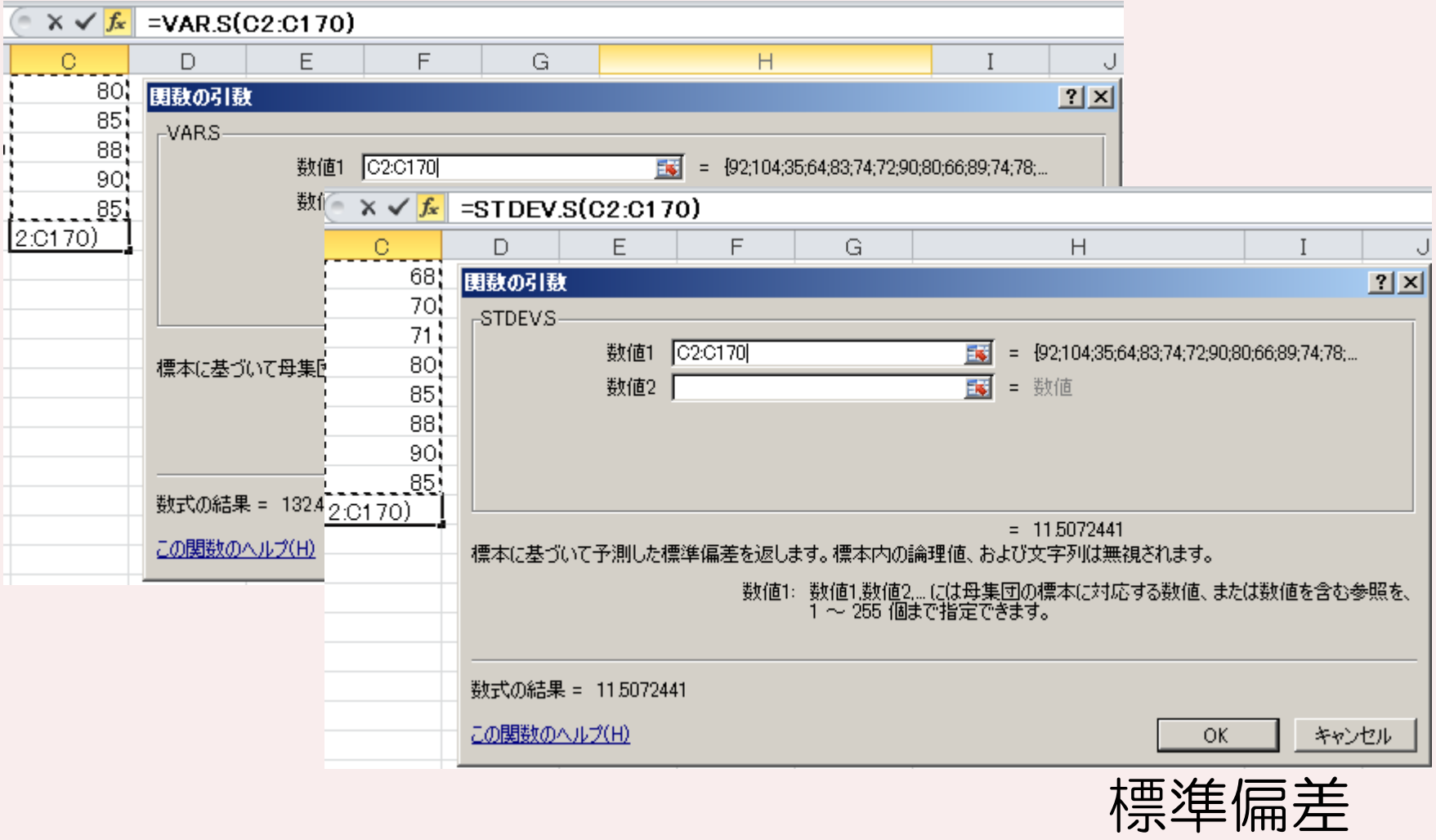
$$\hat{\sigma}_x = s_x = \sqrt{s_x^2}$$

不偏分散

- 母分散と同じ式で標本分散を計算すると，分散を過小推定する（$\bar{x}$ は分散を最小にするμ_xの推定値である）ため補正された不偏分散を用いる．尚，不偏分散を標本分散と呼ぶこともあるので混同に注意すること．
- 従って，不偏分散の式を使用する，即ち$n-1$で割るのは，母平均μ_x が未知で，代わりに標本平均$\bar{x}$ を使う場合のみである．たとえ標本の分散であったとしても，何らかの事情で μ_x が既知であるならば，nで割る点に注意が必要である．

不偏分散

分散


=VAR.S(C2:C170)
関数の引数
VARS
数値1
C2:C170
{92;104;35;64;83;74;72;90;80;66;89;74;78;...
=STDEV.S(C2:C170)
関数の引数
STDEVS
数値1
C2:C170
{92;104;35;64;83;74;72;90;80;66;89;74;78;...
数値2
数値
= 11.5072441
標本に基づいて予測した標準偏差を返します。標本内の論理値、および文字列は無視されます。
数値1: 数値1,数値2,... には母集団の標本に対応する数値、または数値を含む参照を、1 ～ 255 個まで指定できます。
数式の結果 = 11.5072441
この関数のヘルプ(H)
OK
キャンセル

標準偏差

実習

- アデレード大学の学生の身長のデータについて，そのデータを母集団とみなした時と，母集団からの標本とみなした時の平均，分散，標準偏差をそれぞれ求めよ．
- 身長は正規分布に従うと仮定する．得られた母平均，母標準偏差の値を用いて累積40%となる値を求めよ．また，その値から母平均を引き，母標準偏差で割って標準化せよ．
- 関数“NORM.S.INV”を用い，標準正規分布における累積40%となる値を求めよ．