第7章　テキストの探索的分析

茨城大学
佐々木研究室
高木真

はじめに

- テキストから収集したデータを統計分析する際に、データの中心を表す値、データの散らばりの度合いを示す値などのデータの要約値を用いて探索的分析をおこなうことは非常に重要

- データを要約する基本的な統計量について説明する。

中心値①

- 中心値
  - データの中心を表す統計量
- 平均
  - データの合計をデータの個数で割った値
  - 算術平均(略して平均)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

x_i : 項目 i が対応する値

$\bar{x}$: 算術平均

中心値②

- 中央値
  - データを小さいものから順に並べたとき、真ん中に位置する値
  - データの個数が偶数だと中央に位置する値2つ

    ⇒2つの値を足して2で割った値が中央値
- 最頻値
  - データの中で最も頻繁に現れる値

バラツキ①

- 分散
  - データの平均を基準としたバラツキ(散らばり)を示す度合

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

- 標準偏差
  - 分散の平方根
  - 標本のデータから母集団の性質を推測する場合は右側の式を用いるのが一般的

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad s^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, s = \sqrt{s^2}$$

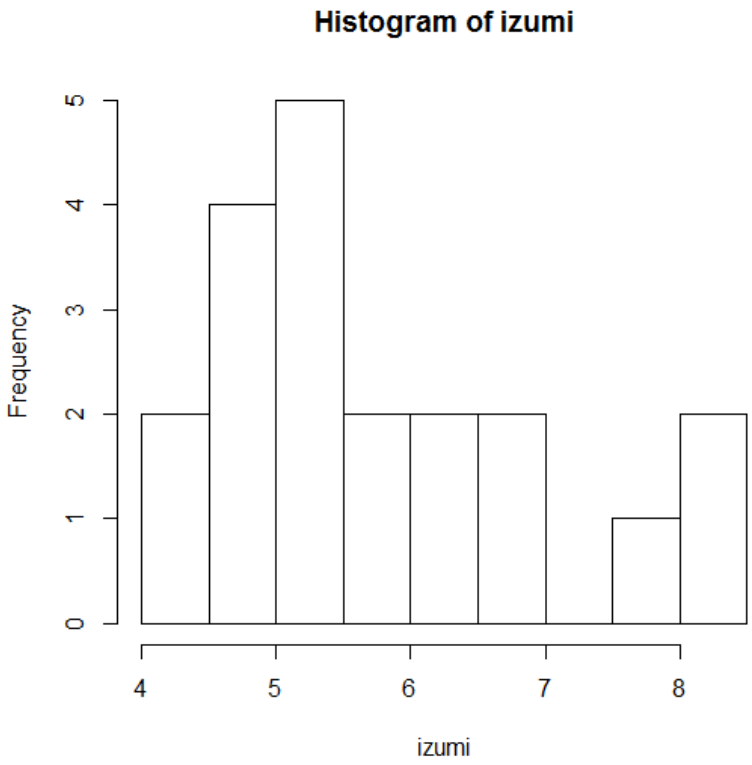
バラツキ②

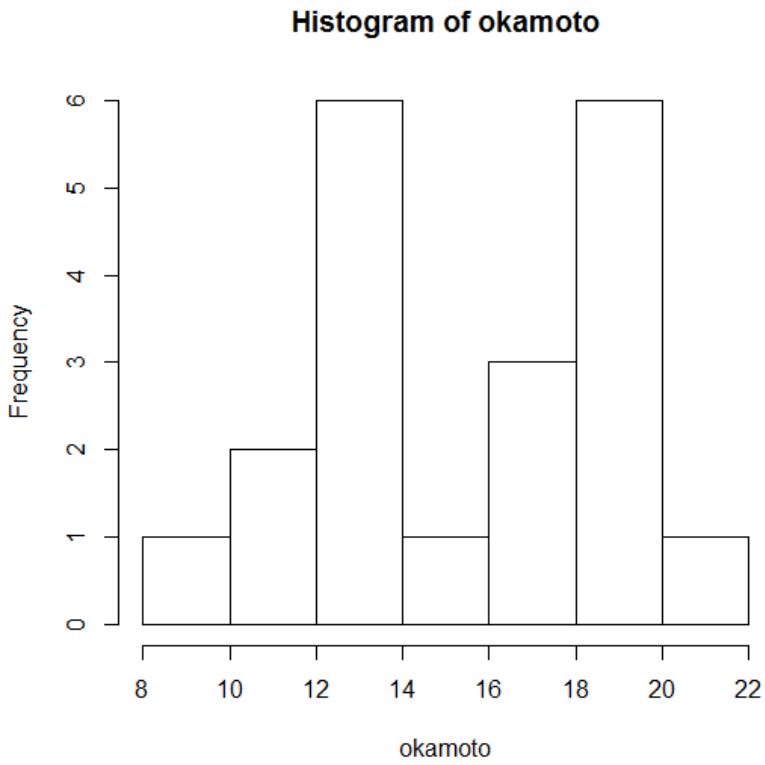
- データの範囲
  - データの最小値と最大値の差を取ったもの
- 分位数
  - データを小さいものから順に並べ、その範囲を等間隔に区切った境値
  - 10等分、4等分(四分位数)が多く用いられている
- 四分位数
  - データの個数が非常に多いときに用いるとデータの全体像が見やすくなる

ヒストグラム①

- ヒストグラム
  - 大量の量的データを分析するときに、データの範囲をいくつかの階級（区間）に分け、階級に属する値を集計して図示する方法
  - 横軸に階級、縦軸にその階級に属する度数を示す統計グラフの一種

泉鏡花と岡本綺堂の作品においての「と」の後に読点を打つ癖についてのヒストグラムの比較

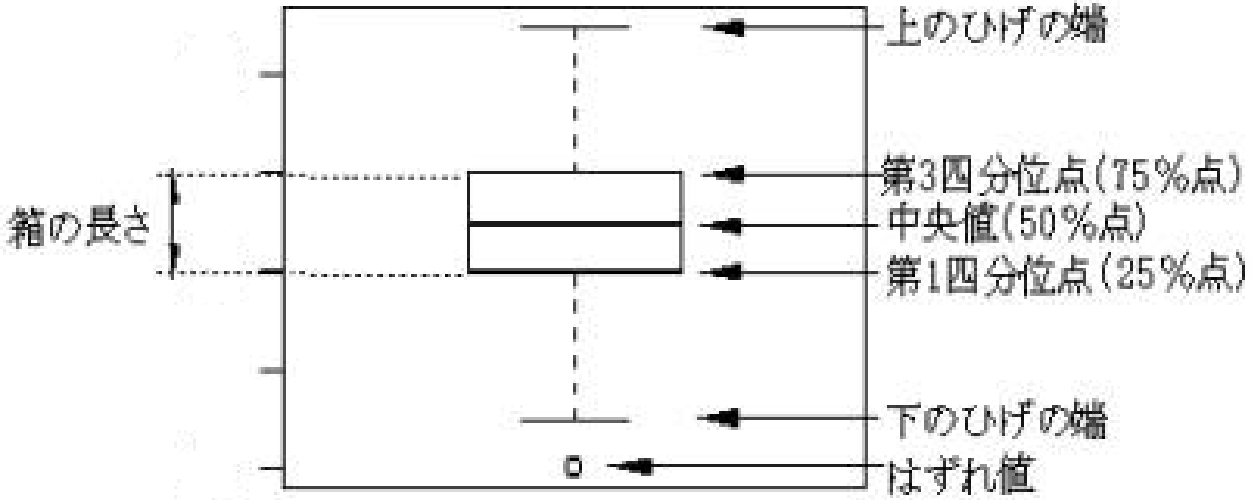




両ヒストグラムの中心がかけ離れていることは、両氏が「と」の後に読点を打つ癖が大きく異なることを意味する。

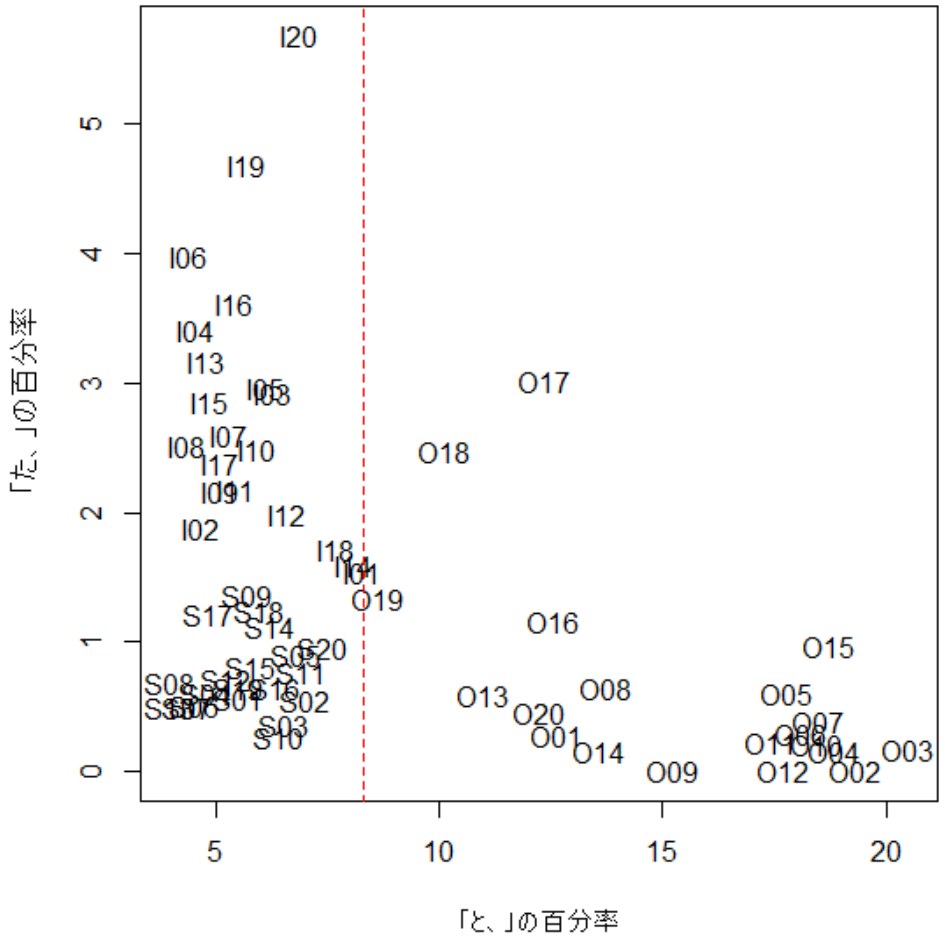
箱ひげ図

- 箱ひげ図
  - データの中心とバラツキを同時に図示する方法
  - 長方形(箱)の両端に直線(ひげ)をつなげたようなグラフ
  - データの中心とバラツキを視覚的に考察できる



散布図

- 散布図
  - ２変数の量的データを図示する方法(ヒストグラム、箱ひげ図は１変数の量的データ)
  - １変数を横軸、もう１変数を縦軸にしてプロットするものが多く用いられる。



「と、」「た、」の使用率の散布図

破線は文章を著者別に分ける境界線

	と、	た、
泉	4.37～8.29	1.54～5.69
岡本	8.62～20.51	0.00～3.02
島崎	3.97～7.40	0.27～1.36

書き手ごとの変数の区間

このような情報を用いると

もし「と、」≧8.45であれば岡本

などの書き手を識別するルールを構築できる。

探索的変数の選択

- 「と、」や「た、」のような著者の特徴が顕著な変数を、多くの変数の中からどのように見つけるかが問題
- 変数が少ない⇒全部の変数について箱ひげ図
- 変数が多い⇒膨大な労力が必要

- データのバラツキについて考える！

- ある変数における書き手の特徴が顕著に表れると、その変数における
  書き手の間(群間)のバラツキは大きく
  書き手ごと(群内)のバラツキは小さく
  なるはずである

偏差の平方和を用いて、どの変数に書き手の特徴が顕著に現れているかについて分析を行う

所属する群が既知であるデータ

$$x_1^a, x_2^a, \cdots, x_{m_a}^a, x_1^b, x_2^b, \cdots x_{m_b}^b, x_1^z, x_2^z, \cdots x_{m_z}^z$$

x_i^g : g 群の中の i 番目のデータ

$$SS = SS_B + SS_W$$

SS : 総偏差の平方和

SS_B : 群間(著者間)の偏差の平方和

SS_W : 郡内(著者別)の偏差の平方和

$$SS = \sum (x_i - \bar{x})^2, \bar{x} = \frac{1}{m}\sum x_i$$

$$SS_W = \sum_{g=a}^{z}\sum_{i=1}^{m_g}(x_i^g - \overline{x^g})^2, \overline{x^g} = \frac{1}{m_g}\sum_{i=1}^{m_g} x_i^g$$

$\overline{x^g}$: 群 g 内の平均

m_g : g 群の個体数

SS_B / SS_W が大きいほど書き手の特徴が顕著

※各群内の分散が等しくないことでこの値が有効じゃない場合がある⇒Gini分散指標(第14章)