統計学　第 4 回
「データを1つの値にまとめる」

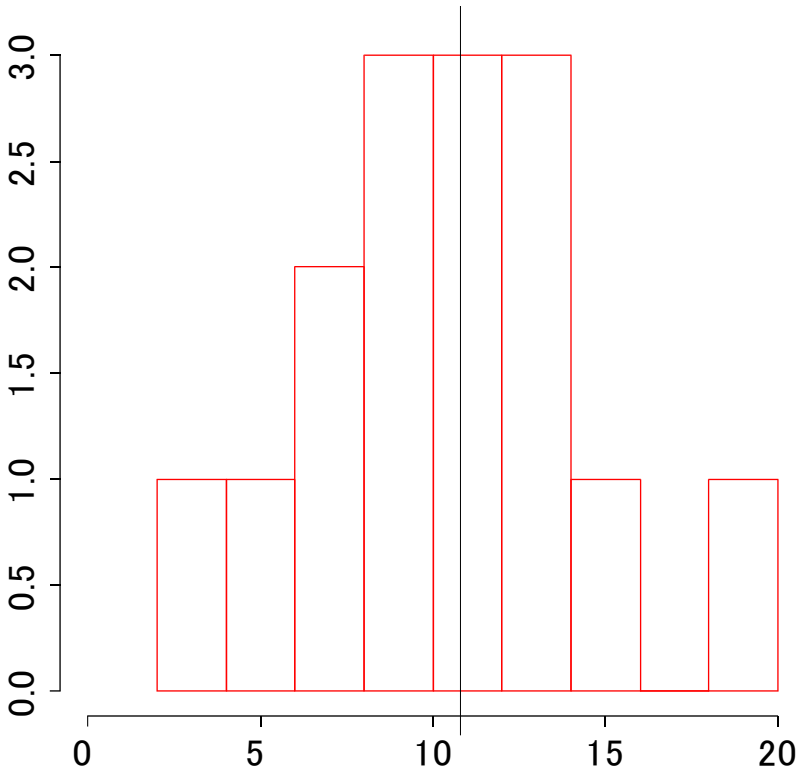
中澤　港（なかざわ　みなと）
内線 1453，E-mail: minato@ypu.jp
http://phi.ypu.jp/stat.html

代表値とは？

- データ全体の情報を1つの値に集約して示すもの。
- 分布の中心（位置）の指標とばらつきの指標がある。

代表値の有効性を示す例

10.5, 11, 12.7, 14, 9, 9.2, 8, 3.5, 19.3, 7.6, 5.4, 11.2, 13.4, 15.3, 10
と数字が並んでいても，だいたいどれくらいかがわかりにくい。
しかし，これをヒストグラムに図示すると，右図のようにわかりやすい。
しかし，たくさんの種類の情報を並べてみるときなど，情報量が多すぎて見通しが悪くなる欠点もある。その解決のためには，1つの値にできると便利である。つまり，
▼ 分布の中心がどのあたりか（目分量だと右図の縦棒辺り），
▼ 分布のばらつき（広がり）がどのくらいか
という情報が役に立つ。前者の代表例が平均値で，この例では約 10.7 である。



この図を描くR のプログラムは，x にデータを入れてから
hist(x,breaks=8,xlim=c(0,20))
平均値を求めるには
mean(x)

分布の中心（位置）の指標

- 英語では central tendency という
- 狭義の代表値
- 平均値，中央値，最頻値がよく使われる。
- 他に幾何平均，調和平均などがある。

平均値 (mean)

- データの総和をデータ数で割った値。物理的に考えれば重心と等しい。
- データ全部を使うが，極端な外れ値や歪んだ分布のときに代表性が弱くなる。
- 間隔尺度または比尺度でないと計算する意味がない。
- （例題）{7, 10, 11,13,18} の平均値は？

中央値 (median)

- データを大小関係に従って順番に並べたときの真中の値。
- 順序尺度以上の尺度で計算できる。
- 外れ値や歪んだ分布に強い。
- （例題）{7, 10, 11,13,18} の中央値は？
- {7,10,11,13,51} の平均値と中央値は？

最頻値 (mode)

- 頻度の最も高い値。
- 名義尺度でも得られる。

分布の中心（位置）のその他の指標

- 幾何平均 (geometric mean) : 積の累乗根。データの分布が対数正規分布に近い場合によく使われる。実際にはデータを対数変換してから平均値を出し，それを元に戻すと幾何平均になる。
- 調和平均 (harmonic mean) : 逆数の平均値の逆数。0 を含むデータには使えないが，外れ値には強い。
- （例題）{7, 10, 11,13,18} の幾何平均と調和平均は？

分布のばらつき (variability) の指標

- 範囲，四分位範囲，四分位偏差，平均偏差，分散，不偏分散，標準偏差，不偏標準偏差といったものがある。
- 他に変動係数 (Coefficient of Variation; CV ＝標準偏差を平均値で割った値）も，相対的なばらつきの尺度としてよく使われる。

範囲 (range)

- 単純に最大値と最小値の差をとったもの。
- 外れ値の影響を受けやすい。
- （例題）{7, 10, 11,13,18} の範囲は？

四分位範囲 (IQR) と四分位偏差 (SIQR)

- 順番に並べたデータを４つに分割して，上から１／４の点と下から１／４の点の差をとったものが四分位範囲，その半分が四分位偏差である。
- 外れ値や分布の歪みの影響を受けにくいという利点をもつ。
- 検定を順位検定や順位和検定など，ノンパラメトリックな（分布を仮定しない）方法でやる場合は，分布が歪んでいたり外れ値がある場合なのだから，代表値も平均と標準偏差ではなく，中央値と四分位偏差で出すべき。
- （例題）{7, 10, 11,13,18} の四分位範囲と四分位偏差は？

平均偏差 (mean deviation)

- 平均値と各データの差の絶対値の総和をデータ数で割った値
- 全データを使う
- １つの外れ値の影響は受けにくい
- 絶対値を使うために，統計量として，他の統計量と数学的な関連をもたない
- 標本統計量から母集団統計量を推定するのには使えない。
- （例題）{7, 10, 11,13,18} の平均偏差は？

分散 (variance) と 不偏分散 (unbiased variance)

- 分散は，平均と各データの差の二乗の和をデータ数で割った値。
- 不偏分散は，データ数で割る代わりにデータ数から1を引いた「自由度」で割ったもの。標本データから母集団の分散を推定するときはこちらを用いる。
- （例題）{7, 10, 11,13,18} の不偏分散は？

標準偏差 (SD) と不偏標準偏差

- 標準偏差は分散の平方根，不偏標準偏差は不偏分散の平方根である。
- もっとも普通に使われるばらつきの指標
- 通常の論文や本や新聞などで発表されている場合は，標本調査で得たデータから母集団統計量を推定しているのが普通なので，わざわざ不偏と書かれていなくても不偏統計量が使われている場合が多い。つまり，全数調査でない限り，SD といえば不偏標準偏差のことを意味していると考えてよい。
- （例題）{7, 10, 11,13,18} の不偏標準偏差は？