R による計量分析：データ解析と可視化

第 3 回　R の基礎とデータ操作・管理

伊藤 岳

富山大学 経済学部 2017 年度後期

Email: gito@eco.u-toyama.ac.jp

October 23, 2017

Agenda

1. 計量分析とプログラミング
2. データと統計量
3. 中心極限定理
4. R の基本的操作と注意点
5. RStudio での演習

休講予定更新

次の日程は休講

- 10/30 (Mon.)
- 12/11 (Mon.) ← New!
- 1/9 (Tue.) ← 月曜講義日
- 補講については講義の進捗具合をみながら，今後アナウンスする
- たぶん補講設定 ← New!
  - 補講時期は，特にリクエストがなければ年明け

講義で扱うデータ解析・可視化の概念と手法

データの解析

- 回帰分析 (regression analysis)：線形回帰 (OLS) と一般化線形モデル (GLM)
  - 相関関係と因果関係
  - 内生変数と外生変数
  - 回帰分析の解釈と限界効果
- 確率論の基礎
- 推測統計 (inferential statistics)
  $\Longrightarrow$ 本格的には講義の後半で扱うが，今日も導入的な内容を扱う (講義後半は課題文献が増えるので注意)

データの可視化 (とデータの「準備」)

- 可視化の目的：次元の縮約・要約
- 基礎的な可視化：ヒストグラムや散布図
- 発展的な可視化：ネットワーク，地図
- (データの自動取得，加工)

講義で扱うデータ解析・可視化の概念と手法

本当のような「怪しい」話

- 「広告費を倍増した結果，アイスクリームの売り上げが 50%伸びた」
- 「新社長の改革の成果によって，株価が 30%上昇した」
- 「政府の補助金政策の効果で，地域経済が活性化した」
- 「マンションの高層階に住むと，妊娠しにくくなる」
  - 出所：伊藤 公一朗「Google 検索の『青色』に隠された最強の分析力：世界の勝ち組企業はビッグデータをこう使う」(`http://toyokeizai.net/articles/-/171160?display=b`)
  - 上記の例がなぜ「怪しい」かを平易に解説
  - 興味があれば，副読本を参照：伊藤 公一朗．2017.『データ分析の力：因果関係に迫る思考法』光文社

「きちんと評価する」ためには

- 計量分析の手法・発想が有用
- 再現可能な形で評価を提示するには，R (プログラミング) が有用

研究の再現可能性

「科学」性とプログラミング

- 経済学であれ政治学であれ社会学であれ，社会「科学」であるためには，再現可能でなくてはならない
- エクセルなど「ポチポチ」するようなソフトウェアの利用は，こうした条件を満たすことを妨げる
- 「ポチポチ」するソフトウェアの利用は，自分自身のためにもならない
  - 「一ヶ月前に行った解析」を「今」再現できる？

研究の再現可能性を担保するために

必要な記録

- データの出所，取得方法
- データの加工方法
- データの解析方法
- 解析結果の解釈
- これらすべてをコードで記述し，データセットとあわせて (原則) 公開する

研究成果の提出に際しては

- コードとデータセットを公開することが (原則) 必須
- この講義の課題でも，コード提出を求める (ゲーム理論のテストで証明を書くのと同じ)

データとは何か

辞書的定義

- 材料，資料，論拠という意味の datum の複数形．コンピュータ用語として，情報を作成するために必要な資料の意味に使われる．コンピュータに入力する記号，数字，文字のことで，それ自体，単なる事実にすぎず，コンピュータにより，一定のプログラムに従って処理されて，特定の目的に役立つ情報を生む (『ブリタニカ国際大百科事典 小項目事典』)
- (1) 判断や立論のもとになる資料・情報・事実．「— を集める」．(2) コンピューターの処理の対象となる事実．状態・条件などを表す数値・文字・記号 (『大辞林 (第三版)』)
- 出所：「コトバンク」(`https://kotobank.jp`)

- 調査・実験・観察によって収集された情報の集合のこと
- (データを可視化・解析することで，) 判断・立論の基準にもなる
- 基本的には，定量的データ (連続／量的変数) と定性的データ (質的／離散変数) に分けられる
  - 定量的データの例：人口，GDP
  - 定性的データの例：同盟国，国名

観察単位と変数

観察単位

- 観察単位 (unit of observation)：観察の対象，記録の単位
  - 例：国家，自治体，個人
  - 解析の際は分析単位 (unit of analysis) とも

変数

- 変数 (variable)：観察単位で計測・記録され，観察単位によって変化する値
  - 例：GDP, 人口，収入，性別
  - 2 つ以上の値をとる
  - 質的なものでも量的なものでも可：GDP でも性別でもよい
  - 同じ値が複数あってもよいが，定数 (constant) ではない
  - 例：「日本人のデータ」なら，「性別」「身長」は変数，「国籍」は定数 (二重国籍以外)

変数と尺度

- 質的変数と量的変数は，さらに 4 つの尺度水準に分類される
  1. 名義尺度 (nominal scale): 整理番号のような尺度.「2 つの値が同じか否か」(異同) のみを示す尺度
     - 例：性別，国籍，出身地
  2. 順序尺度 (ordinal scale): 2 つの値の間の大小関係を示す尺度．2 つの値の異同と順序を示す尺度
     - 例：順位．1 位と 2 位が違うこと (異同) と，1 位の方が 2 位より上の順位ということはわかる (順序) が，その間の「差の大きさ」は分からない
  3. 間隔尺度 (interval scale): 目盛が等間隔になっていて，その間隔 (和と差) のみに意味がある尺度
     - 例：摂氏温度，湿度．5 度から 10 度までの気温 5 度の差には意味があるが，比率 (2 倍) には意味がない (0 は「ひとつの状態」)
  4. 比例尺度 (ratio scale): 数値の差とともに数値の比にも意味がある尺度．絶対原点 (0) がある尺度 (0 は「何もない状態」)
     - 例：身長，体重，速度．体重 50kg と 100kg の差は，50kg の差とも，2 倍 (比率) ともいえる
- 情報量：比例 $>$ 間隔 $>$ 順序 $>$ 名義
- 一般に,「情報量を減らす」ことは可能，逆は不可能

データを特徴付ける：統計量

統計量 (statistic)

- あるデータ／変数の特徴を要約する値のこと
- 一定の (統計学的な) 方法・関数 (アルゴリズム) を適用して得る
- (基本／要約統計量には) 大別して，中心的傾向 (代表値) を示す統計量と，ばらつき (散布度) を示す統計量がある
  - 例：算術平均，分散
- 質的変数と量的変数によって，用いる統計量は異なる

統計量の例：平均値，中央値，最頻値

算術／相加平均 (mean, average)

n 個のデータ (実数) $x = (x_1, x_2, \ldots, x_n)$ の平均値 $\bar{x}$ は,

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

中央値 (median)

n 個のデータ x の中央値 m は，次の不等式を満たす実数値である.

$$\int_{-\infty}^{m} \mathrm{d}f(x) \geq \frac{1}{2} \quad \text{and} \quad \int_{m}^{\infty} \mathrm{d}f(x) \geq \frac{1}{2} \tag{2}$$

- 言い換えれば，m は「データの真ん中の値」
- n が偶数なら,「真ん中前後の 2 つの値」の算術平均が中央値になる

統計量の例：平均値，中央値，最頻値

最頻値 (mode)

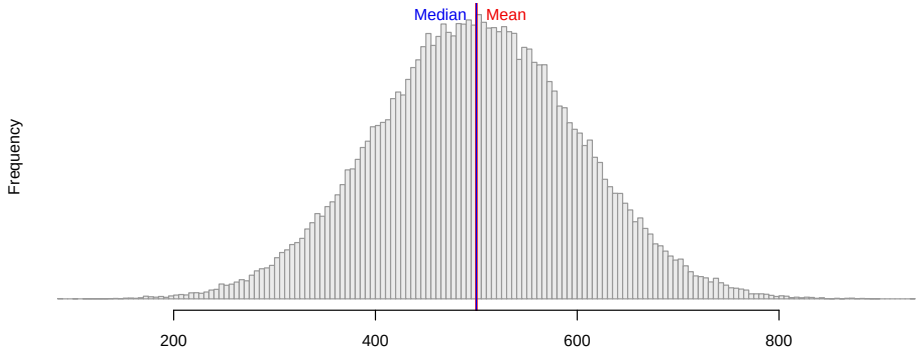
あるデータ，確率分布において，(度数分布で) 最も頻度の高い値のこと．たとえば，$x = (1, 1, 1, 1, 1, 2, 3, 4, 5, 6)$ であれば，1 が最頻値となる

3 つの統計量の特性

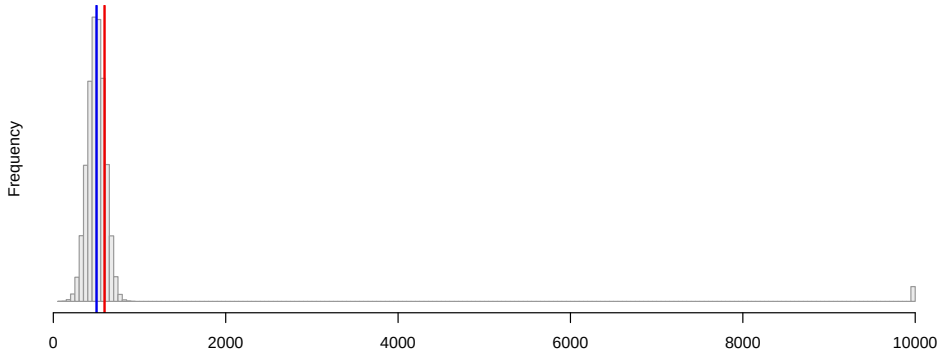
- 平均値は，外れ値 (outlier) の影響を受けやすい．他方，中央値は受けにくい
- 左右対称の分布 (e.g., 正規分布) では，平均値と中央値は一致する
  - 平均値と中央値のイメージは，次スライド
- 平均値は「ヒストグラムの重心」を，「中央値はヒストグラムの真ん中」を，最頻値は「ヒストグラムのピーク」を示す
- ただし，最頻値は一意に定まらないこともある (e.g., 一様分布や連続量)

平均値と中央値

平均 500, 標準偏差 100, $n = 10,000$ の正規分布：$(\bar{x} = 500, m = 500)$



上の正規分布に，10^4 を 100 個加えた分布 $(\bar{x} = 594, m = 501)$



- $100/10,000 = 1/100$ の外れ値の追加に対して平均値は敏感に反応し，中央値はあまり変化しない
- 平均値は外れ値の影響を受けやすく，中央値は受けにくい (外れ値に対して頑健 robust)

統計量の例：分散と四分位範囲 (IQR)

不偏分散 (unbiased variance)

n 個のデータ $x = (x_1, x_2, \ldots, x_n)$ の平均値を $\bar{x}$ とするとき，不偏分散 σ_x^2 は

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{3}$$

である．また，σ_x^2 の平方根 σ_x を標準偏差 (standard deviation, sd) と呼ぶ．

- 直感的には，「平均値からのズレ」(を二乗した値) の平均値
- 正の方向にズレることもあれば，負の方向にズレることもあるので，二乗する
- (3) 式より，変数の (不偏) 分散は必ず正の値をとる
- 分母の $n-1$ を n としたものを，(母) 分散と呼ぶ
- 分散と標準偏差
  - 分散は (差を) 二乗しているので，元のデータ x や平均 $\bar{x}$ と直接比較できない
  - 標準偏差は，平方根をとることで「次元を戻している」

統計量の例：分散と四分位範囲 (IQR)

四分位範囲 (inter-quartile range, IQR)

昇順 (小さい順) に並べた (1 次元の) n 個のデータを $x = (x_1, x_2, \ldots, x_n)$ とするとき，x の四分位数とは，x を 4 つの個数の等しいグループに等分したとき，各グループの境界線となる値をいう．四分位範囲 IQR とは，第 3 四分位数 $Q_{3/4}$ (upper quartile) と第 1 四分位数 $Q_{1/4}$ (lower quartile) の差 $Q_{3/4} - Q_{1/4}$ を指す

- 分散と同様に，四分位範囲はデータのばらつきを示す統計量
- 中央値 $m = Q_{2/4} = Q_{1/2}$ であり，$Q_{0/4}$ は最小値，$Q_{4/4}$ は最大値を示す
- 平均と標準偏差 (分散)，中央値と四分位範囲はセットで使う
- 四分位範囲 IQR には，中央値周辺に並ぶ約 50%のデータが含まれる
- $[Q_{1/4} - 1.5\mathrm{IQR}, Q_{3/4} + 1.5\mathrm{IQR}]$ の区間に含まれない値を，外れ値 (outlier) とみなす
  - 「箱ひげ図 (box-and-whisker plot)」で可視化する (次回描いてみる)
  - なお，$q/10$ 分位数は，第 q 十分位数と呼ぶ

標本と母集団

- 母集団 (population)：研究において，興味がある／理解したい対象全体のこと．一般的に，母集団を直接観察することはできない
  - 詳しくは (確率分布とあわせて) 講義の中盤以降で扱う
  - 教科書 (星野・田中本) の通り,「母集団」の捉え方には 2 つの種類がある
  - データ生成過程 (data generating process) とも密接に関わる
- 標本 (sample)：データとして観察された，母集団の一部
  - 母集団から標本を得ることを，標本抽出 (sampling) という
- 通常，データとして観察できる (手に入る) のは母集団ではなく標本
- 統計的推定 (statistical inference) の目的 (とありがたみ)：一定の誤差 (error) を許容した上で，手元にある標本 (部分) から母集団 (全体) を知ること

標本数と標本サイズ

- 標本サイズ (sample size): 標本の大きさのこと．ある標本に含まれる (変数の) 観察値の数のこと．一般的に N といわれる数のこと
- 標本数 (number of samples): 標本 (群) の数のこと．観測値の集合の数のこと
  - 正しい用法：10 回の実験の各回で得られた標本サイズはそれぞれ 20
  - 正しい用法：上の実験の繰り返しから得られた標本数は 10
  - 誤った用法：各回で得られた標本数はそれぞれ 20
  - 例：たとえば，日本人 1,500 人のデータとアメリカ人 2,000 人のデータがあったとき，標本数は 2，標本サイズはそれぞれ 1,500 (日本人の標本) と 2,000 (アメリカ人の標本)

母数 (パラメータ)

- 母数 (parameter) の意味にも注意
  - 分母ではない！
  - 母集団・分布を特徴付けるパラメータ (parameter) のこと．講義で扱う推定手法で，推定したい値のこと (e.g.，「真の」回帰係数)
  - 誤った用法：「今春卒業した大学生の就職率について，文部科学省は 1 日，『就職氷河期』と呼ばれた 2000 年春を下回り，過去最低の 91.0%だったと発表している．文科省の調査は母数を『就職希望者」としているのに対し，今回の調査は卒業生全体を母数に取っている」「今春大卒 2 割，進路未定 学部間差，最大 5 倍」『朝日新聞』「ひらく 日本の大学」2011 年度調査結果報告 (`http://www.asahi.com/edu/hiraku/hiraku2011/article01.html`)
- 統計的推定の目的 (再掲)：一定の誤差 (後々出てくる標準誤差 standard error) を許容した上で，パラメータを推定すること
  - 例 1：標本として得られた有権者の情報 (政権支持率) を用いて，日本の有権者全体の政権支持率を推定する
  - 例 2：標本として得られた国家の情報を用いて，経済状況が内戦の発生確率に与える影響を推定する

中心極限定理

中心極限定理 (Central Limit Theorem, CLT)

ある同一の分布から独立に得られた大きさ (標本サイズ) n の標本 $X_1, X_2, \ldots, X_n$ の平均を $\overline{X}_n$, 分散を σ_X^2 とする．また，この分布に従う確率変数 X の期待値を $\mathrm{E}[X]$ とする．標本の大きさ n が大きくなるにつれ，以下の統計量 Z_n ($\overline{X}_n - \mathrm{E}[X]$ は標本平均と母平均の誤差) は，平均 0, 標準偏差 1 の正規分布 (標準正規分布) $\mathcal{N}(0, 1)$ に近づく (弱収束する).

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mathrm{E}[X])}{\sqrt{\sigma_X^2}} = \frac{\sqrt{n}(\overline{X}_n - \mathrm{E}[X])}{\sigma_X} \tag{4}$$

- $\overline{X}_n - \mathrm{E}[X]$ が $\mathcal{N}(0, \sigma_X^2/n)$ に近づく，でも同じ意味 (正規分布の標準化)
- 標本サイズ n が十分に大きければ，元の分布どんなものであれ誤差 $\overline{X}_n - \mathrm{E}[X]$ の分布は正規分布に近くなる
  - ただし，元の分布に平均値と分散が存在すれば
  - 平均 μ, 標準偏差 σ (分散 σ^2) の正規分布を，一般的に $\mathcal{N}(\mu, \sigma^2)$ と書く

中心極限定理

中心極限定理の帰結 (正規分布が出てくると何が嬉しいのか)

ある同一の分布から独立に得られた大きさ (標本サイズ) n の標本 $X_1, X_2, \ldots, X_n$ の平均を $\overline{X}_n$, 分散を σ_X^2 とする．また，この分布に従う確率変数 X の期待値を $\mathrm{E}[X]$ とする (ここまではさっきと同じ)．標本の大きさ n が大きいとき，以下の不等式が，95% の確率で近似的に成立する．

$$\overline{X}_n - 1.96\sqrt{\sigma_X^2/n} \leq \mathrm{E}[X] \leq \overline{X}_n + 1.96\sqrt{\sigma_X^2/n} \tag{5}$$

標本平均 $\overline{X}_n$ は，正規分布 $\mathcal{N}(\mathrm{E}[X], \sigma_X^2/n)$ に従う

- 母数を 95%の精度で含む信頼区間を計算できる！ (言い方に注意)
- 統計的推定・検定の基礎になる定理 $=$ 正規分布が「大事にされる」理由

信頼区間

中心極限定理と信頼区間

- (5) 式の区間を $\mathrm{E}[X]$ の 95% 信頼区間 (confidence interval, CI) と呼ぶ
- 標準誤差 (standard error, SE): $\sqrt{\sigma_X^2/n} = \sigma_X/\sqrt{n}$
  - (不偏) 標準偏差 σ_X を $\sqrt{n}$ (標本サイズ n の平方根) で割った値
  - つまり，95% CI は $[\overline{X}_n - 1.96\mathrm{SE}, \overline{X}_n + 1.96\mathrm{SE}]$
- 95% 信頼区間を使って，母数を推定できる
  - 統計的推定で問題になるのは，標本統計量と母数の誤差
  - 誤差 $(\overline{X}_n - \mathrm{E}[X])$ が正規分布なら，その性質を利用できる
  - n が大きくない場合は，(正規分布ではなく) t 分布で近似する
  - t 分布については，「なぜ 1.96 が出てくるのか」とあわせて次回以降で説明

信頼区間 (補足)

- $\alpha\%$ 信頼区間の α を信頼係数 (confidence coefficient) と呼ぶ
  - 慣習的な信頼係数：$95\%, 90\%, 99\%$ (原理的には，$94\%, 96\%$, etc. でもよい)
  - それぞれ，5% 有意，10% 有意，1% 有意に対応する
  - つまり，$p < 0.05, p < 0.1, p < 0.01$ (「第一種過誤 (Type I/α error)」(偽陽性) をおかす危険率が $5\%, 10\%, 1\%$)
    - 「第一種過誤 (Type I/α error)」：帰無仮説 H_0 (e.g.,「病気ではない」) が真にもかかわらず，H_0 を棄却してしまう誤り
- 区間推定 (interval estimation)：信頼区間を使った,「幅」を持たせた推定
  - ざっくりした例：「母集団から 100 回標本をとってきて，各々の標本平均 (あるいは，他の統計量) から母平均の 95%信頼区間を求めるという作業を繰り返したとき，95 回については 95%信頼区間の中に母平均が含まれる」
  - 「幅」を持たせる区間推定に対して，1 つの値で母数を推定することを点推定 (point estimation) と呼ぶ

標準偏差と標準誤差

- $\mathrm{SD} = \sigma_X$, $\mathrm{SE} = \sigma_X/\sqrt{n}$ なのだから，常に $\mathrm{SD} > \mathrm{SE}$
- 標本サイズ n が分母にあるのだから $(n \geq 2)$，n を大きくすればするほど，標準誤差は小さくなる
- 標本サイズ n が大きいほど，母数と標本統計量の誤差を正確に推定できる (信頼区間の幅を狭くできる)
  - n が大きいほど，標準誤差 $\mathrm{SE} = \sigma_X/\sqrt{n}$ が小さくなる
  - SE が小さくなれば，95% 信頼区間 $[\overline{X}_n - 1.96\mathrm{SE}, \overline{X}_n + 1.96\mathrm{SE}]$ も狭くなる

信頼区間の解釈と注意

何の信頼度合いか

α%信頼区間：1 つの標本から得られた区間の信頼度ではなく，区間を求める手続きの信頼度を示す

誤：「1 つの標本から得た α%信頼区間の中に，母数が含まれる確率は α%」

ざっくりした例 (再掲) と解釈

「母集団から 100 回標本をとってきて，各々の標本平均 (あるいは，他の統計量) から母平均の 95%信頼区間を求めるという作業を繰り返したとき，95 回については 95%信頼区間の中に母平均が含まれる」

- 標本数が 100 だから，100 個の 95%信頼区間が得られる
- 100 個の 95%信頼区間のうち，95 個は母数を含み，5 個は母数を含まない
- 「各々の 95%信頼区間が母数を含む確率」は，0 か 1
- 手元の標本から得た信頼区間は,「たまたま」母平均を含まないかも知れない

ファイルパスと拡張子

ファイルパス

- (file) path: ファイルやディレクトリの「場所」，そこに至る経路
- 直感的には,「ウェブサイトの URL の，PC 内部版」
  - URL の例：`http://cfes-project.eco.u-toyama.ac.jp/education/education_2017/r_2017/`
  - 例：伊藤の環境のデスクトップにある “sample” というディレクトリ (フォルダ) の path は `/Users/Gaku/Desktop/sample`
  - 例：伊藤の環境のデスクトップにある “sample.csv” というファイルの path は `/Users/Gaku/Desktop/sample.csv`
  - path の取得方法：分からなければ,「自分の OS (Win 10 とか) ファイル パス 取得」等で Google!

拡張子

- 拡張子：ファイル名の最後の “.” 以下の部分のこと
  - 上の例のように，ファイルのパスには拡張子まで入る
  - 例：“sample.csv” なら “.csv,” “sample.xls” なら “.xls”
  - ファイル名に拡張子が表示されていなければ，「自分の OS (Win 10 とか) ファイルパス 取得」等で Google!

R と注意点：言語環境

- 日本語厳禁，英語推奨
- (R 言語や path, encoding 等を深く理解していない限り) R に読み込むファイル名やパスには日本語を絶対に含めてはならない
  - ファイルやコードを保存する際の名前に日本語を入れがちなので，特に注意
  - “/” や “:” のような一部の例外を除いて，半角文字／英数なら問題ない
  - 英語を使うと分かりにくい場合は，ローマ字表記にするなり，メモを残しておく
- R も英語環境を勧める (たぶんその方が質問対応もスムーズ)
- ユーザ名を日本語にしている場合には，(1) ユーザ名をアルファベット表記に変更するか，(2) ルートディレクトリ (最上位のディレクトリ／フォルダ) の直下に，アルファベット名の新たなディレクトリを作成することを推奨
  - Mac なら Macintosh HD (パスは “/” だけ)，Windows なら C ドライブ直下

R と注意点：言語環境

- R は，半角文字と全角文字を「違うもの」として認識する
  - たとえば，“a” (半角文字) と “ａ” (全角文字) を「違うもの」として認識する
  - 「日本語厳禁」なら，この点を考えなくてよくなる
- また，大文字と小文字も「違うもの」として認識する
  - たとえば，“a” と “A” を「違うもの」として認識する
  - 大文字と小文字の区別は，R 言語に特殊
- この辺りは，今日の演習時間で実際に体験する

R と注意点：コードの実行とエラー

- コードの実行：R に限らず，プログラミング言語を処理するシステムは，コードを「上から一行一行順番に実行していく」
  - したがって，途中の「処理 A」にエラーが出れば，「処理 A の完了を前提にしている，それ以後の処理」にもすべてエラーが出る
  - 処理 A のエラーを修正して「処理 A の部分だけ」を回し直しても，コード全体は実行されない
  - 修正して「処理 A の部分」と「処理 A に依存する部分全部」を回し直さなければならない
- R のエラー・メッセージ：「何がおかしいか」「どんなエラーか」を教えてくれる
  - エラー・メッセージを Google すれば，簡単に解決策が出てくることも
  - 例：`"Error: object 'x' not found"`
  - このような「オブジェクトが見つからない」というエラーは，(1) 打ち間違いか (2) コードの実行順序を間違えた場合がほとんど
  - 打ち間違いは本人しか気付かない場合もあるので，よく見直すこと

R 言語の基本的発想：オブジェクト (object) が中心

- オブジェクト：「何かしらの情報を保持する，名前をつきの入れ物」「R で作成・操作したもの全般」
- R 言語では，常にオブジェクトを用いてデータや解析結果を管理する
- オブジェクトの単純な例は，下記の “x”

```
1 > x <- 1 + 1
```

- “<-” (ここでは “=” でもよい) は，「代入する」という意味
- R では，統計処理を念頭に，複数の種類 (「型」) のオブジェクトが用意されている (演習資料で体験)
  - よく使う「型」：vector，matrix，data.frame (tibble)，list
- オブジェクトの「型」によって，保持できるデータ・情報や，可能な処理が異なる
- 一旦保持したオブジェクトの操作・加工や，一定の処理 (e.g., 数値の変換や回帰分析) を行なうことで，データを整理・可視化・解析する

```
1 > x2 <- x/2
2 > x2
3 [1] 1
```

R 言語の基本的発想：データにも「型」がある

- R では，上述のデータの「種類／尺度」に対応する形で，データの「型」が複数用意されている
- よく使うデータの「型」：double (実数), integer (整数), logical (論理値), character (文字列), factor (因子)

```
> x_num <- 1 + 1
> x_num
[1] 2
> x_chr <- "2"
> x_chr
[1] "2"
> class(x_num)
[1] "numeric"
> class(x_chr)
[1] "character"
```

R 言語の基本的発想：データにも「型」がある

- データの「型」によって，可能な処理が異なる
  - たとえば，平均値を計算したいとき，計算対象のオブジェクトは実数あるいは整数型のデータを保持していなければエラーになる
  - たとえば，「文字列の (のデータ型を保持するオブジェクト) の平均値」は計算できない
    - 無理やりやろうとしても，下記のようなエラーが出る (5–8 行目)
  - 上の例にある `x_chr` を「文字の 2」ではなく「数字の 2」として扱いたい場合は，データ型を変換する (9–10 行目)

```
1  > num_vec <- c(1, 2, 3, 4, 5, 6)
2  > mean(num_vec)
3  [1] 3.5
4  > chr_vec <- c("1", "2", "3", "4", "5", "6")
5  > mean(chr_vec)
6  [1] NA
7  Warning message:
8  In mean.default(chr_vec) : argument is not numeric or logical: returning NA
9  > mean(as.numeric(chr_vec))
10 [1] 3.5
```

R の操作

ここからは講義資料を使います

1. 講義資料ウェブページ内の「演習資料」にアクセス (URL: `http://cfes-project.eco.u-toyama.ac.jp/education/education_2017/r_2017/rcode_fall2017/`)
2. 「R コード」の下の「2. R の基本的操作」の見出しの下にある,「R 言語の基礎, オブジェクトとその要素へのアクセス」に入る
   - R のコードと日本語の解説付きのページが表示される

次回講義と課題文献

- 次回講義：予定を変更して，データの操作・管理 (予定内容) に加えて，データの可視化にも入る (1 週間前倒し)
- データの可視化を通して，図の作り方と意味，相関関係と因果関係の導入的な話にも入る
- 今週最後に出てきた中心極限定理を，R でのシミュレーションを用いて確かめてみる

必須　星野・田中『R による実証分析』第 1–3, 6 章 (教科書)
推奨　Gelman & Hill. *Data analysis*. Chap. 1–2 (教科書)
推奨　浅野・矢内『Stata による計量政治学』第 5 章 (副読本)
推奨　伊藤『データ分析の力』第 1–2 章 (副読本)
課題　講義資料「R 言語の基礎，オブジェクトとその要素へのアクセス」をよく読み，R におけるオブジェクトやデータ型の扱いを復習しておくこと