経済統計分析 2
データの種類と基本的な変形

今日のおはなし.

- データの種類
  - 測れるもの, 測れないもの
  - 離散データ, 連続データ, 属性（質的）データ
  - グラフを描こう
- 基本的な変形
  - 基準をそろえる
  - 時系列に特有の変形
- 今日のタネ
  - 中村隆英ほか. 1984.『統計入門』東大出版会, 第2章
  - 上田尚一. 2003.『統計の誤用・活用』朝倉書店, 2〜3章

データの整理:一覧表にする

変数名／個体番号	X	Y	Z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
:			
i	x_i	y_i	z_i
:			
n	x_n	y_n	z_n

表現の約束

変数Xに対して,それぞれのデータを$x_1, x_2, \ldots, x_i \ldots, x_n$と表す

時系列データの場合は，$x_1, x_2, \ldots, x_t \ldots, x_T$と書くことも

- 観測値（observation）
  - (x_1, y_1, z_1) の組合せのこと
  - または,それぞれのx_1, y_1, z_1
  - 添え字をつけて各個体を区別
- データの大きさ･サイズ
  - 観測値の数 n
- 変数･変量
  - (X, Y, Z) のような,量的な特性
- 行と列
  - 好み,統計処理ソフトによる
  - MS-Excelは縦長が得意

測れるもの，測れないもの

- 統計処理の対象とするのは基本的に測れるもの
  - 「目に見えないもの」「計測できないもの」は対象にできない
  - 「計測できないもの」の代理変数（proxy）はよく使う
  - 例：国の経済の開放度 = 輸出入額/GDP

- 計測可能でも数値にしにくいものは多い
  - 統計（ソフトの）処理の都合上，数値を当てはめることも
    - コード番号をつける
  - 例：属性（性別，職業，…），手段（交通手段，…），有無
  - 例：傾向（支持する，やや支持する，どちらでもない，…）
    - 数値の大小が意味を持つばあいもある

連続･離散･質的変数

- 連続変数 continuous variable
  - 変数のとりうる値が連続的なもの
  - 数学的な意味での連続性ではないことが多い
  - 整数値しかとらなくても，連続とみなすのが普通：計量データ
- 離散変数 discrete variable
  - 変数のとりうる値がとびとびなもの
  - 個数を数えて得られるデータのことが多い：計数データ
- 質的変数 qualitative variable
  - 調査対象の質的な特性を調べたデータ
  - コード表にしたがって整数を当てはめて扱うことがおおい
  - 整数の値そのものに意味はない
  - 「1-0」で表現するとき，ダミー dummy 変数と呼ぶ

基本的な表, グラフ:横断面1変数

- 度数分布(頻度分布)表, ヒストグラム
  - 質的変数のばあいは, 各値を取る観測値の数をまとめたもの
  - 連続変数のばあいは, ある区間(階級)の値を取る観測値の数
  - 階級値:階級を代表する値. しばしば中位値
- 相対度数
  - 各階級の度数を全体の度数で割ったもの. [0, 1] の実数値.
  - 「率」は相対度数の一種
- 累積度数
  - 低い階級から度数を順に足して得られる度数
  - 累積度数の差をとると度数が得られる
- 累積相対度数
  - 累積度数を全体の度数で割ったもの. [0, 1]の単調増加

基本的な表：横断面1変数

階級	階級値	度数	相対度数	累積度数	累積相対度数
$a_1 \sim b_1$	m_1	f_1	f_1 / n	$F_1 = f_1$	F_1 / F_n
$a_2 \sim b_2$	m_2	f_2	f_2 / n	$F_2 = F_1 + f_2$	F_2 / F_n
⋮	⋮	⋮	⋮	⋮	⋮
$a_j \sim b_j$	m_j	f_j	f_j / n	$F_i = F_{i-1} + f_i$	F_i / F_n
⋮	⋮	⋮	⋮	⋮	⋮
$a_k \sim b_k$	m_k	f_k	f_k / n	$F_k = F_{k-1} + f_k$	$F_k / F_n = 1$
計		n	1	$F_n = F_k$	

- 度数分布／ヒストグラムの作り方
  - 階級の区間を等間隔にとるほうが望ましい
  - 階級の境界か，階級値がきりのいい数値になるように
  - 階級の個数はデータのサイズと要相談
  - グラフ作成時にはオープンエンドや階級幅を反映させる

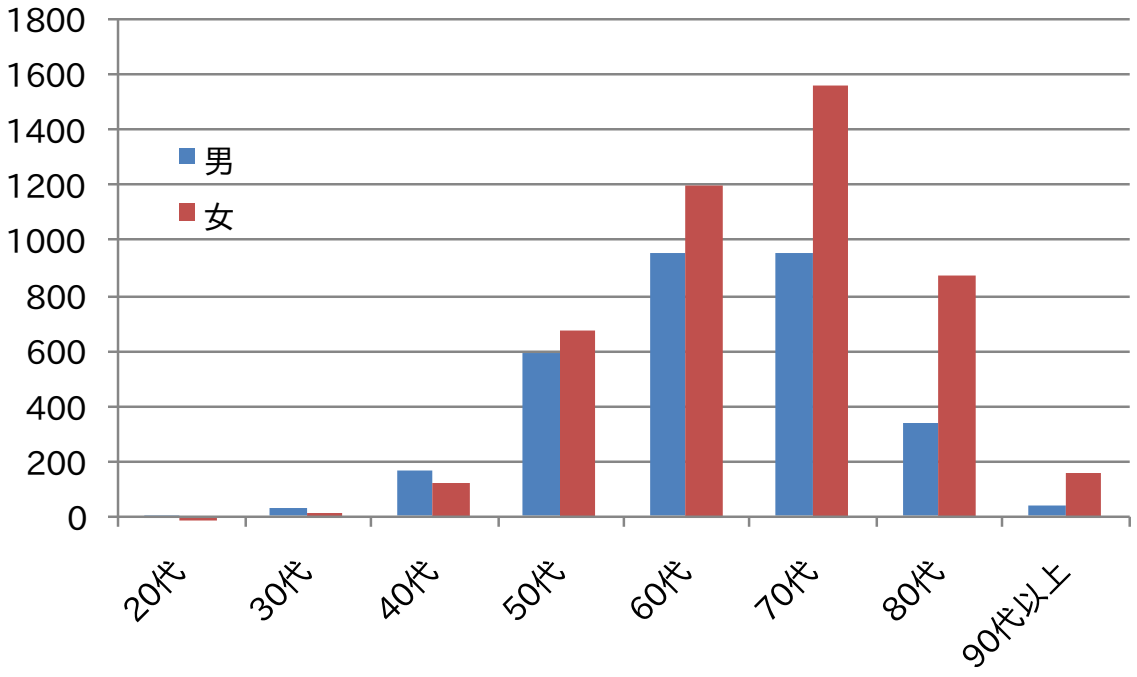
基本的な表, グラフ：横断面2変数(以上)

- 度数分布表（クロス表）
  - 制御したい変数Xの階級ごとの検討したい変数Yの度数分布
    - 例：男女別の賃金分布, 年齢階級別の賃金分布
  - 3変数以上のばあいは, 変数を組み合わせて変数Xを作る
    - 例：男女-年齢階級別の賃金分布：男性20歳代, 女性20歳代, 男性30歳代….
  - ヒストグラム：グラフを2つ描いて並べるか, 棒を並べて描く
- 散布図 scatter plot
  - (x_i, y_i) をX-Y平面上に打点したもの
  - 3変数のばあいは, グラフを2つ描いて並べるか, 変数ごとにマークを変える

基本的な表, グラフ：横断面2変数(以上)

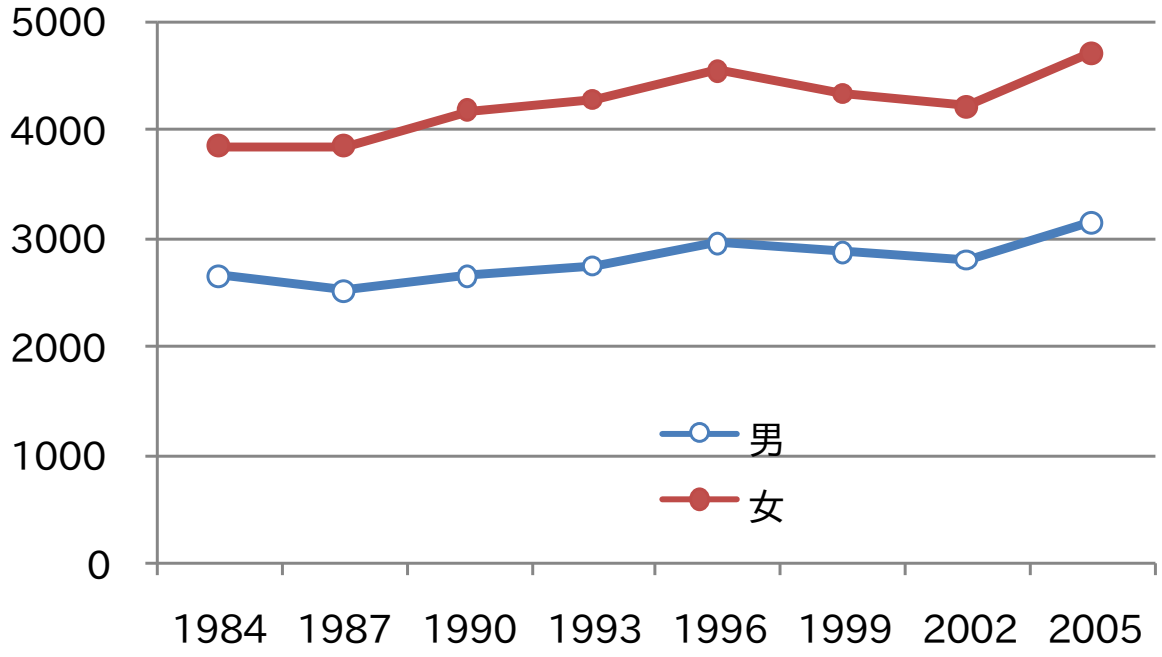
例：本態性（原発性）高血圧（症）（2005年, 千人）

（千人）	20代	30代	40代	50代	60代	70代	80代	90代以上
男	3	34	167	597	958	957	339	39
女	1	20	128	676	1206	1569	882	161



基本的な表, グラフ:時系列

- 一覧表
  - 一定の期間での平均値をとることも
- グラフ
  - 横軸に時間軸をとって折れ線グラフ

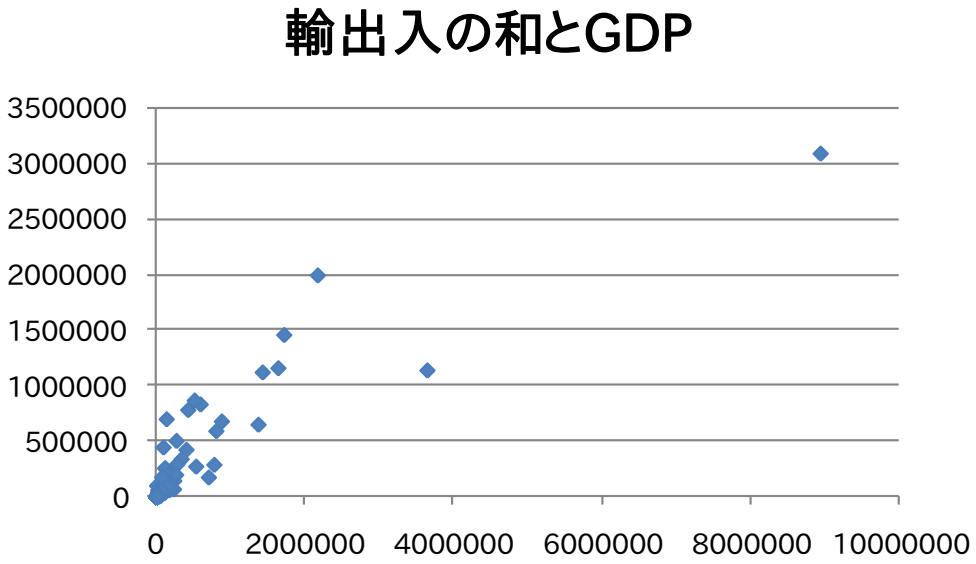


基本的な変形

- 基準化
  - 観測値間の比較を行うため,「本質的」ではなくて明らかな変数の影響を除去するためになんらかの基準化を行う
  - 基準化した変数同士の散布図によって関係性をみよう
- よくある基準
  - 人口一人当たり
  - 面積あたり:人口密度は代表例
  - 経済規模あたり:対GDP比, 対税収比
- どの基準を使うのがよいか, は状況に依存するが
  - その基準化が意味を持つか?
  - 検討の対象となっている変数と, 基準化に使おうとしている変数の散布図を描くと, 直線状になっているか?

基本的な変形：横断面

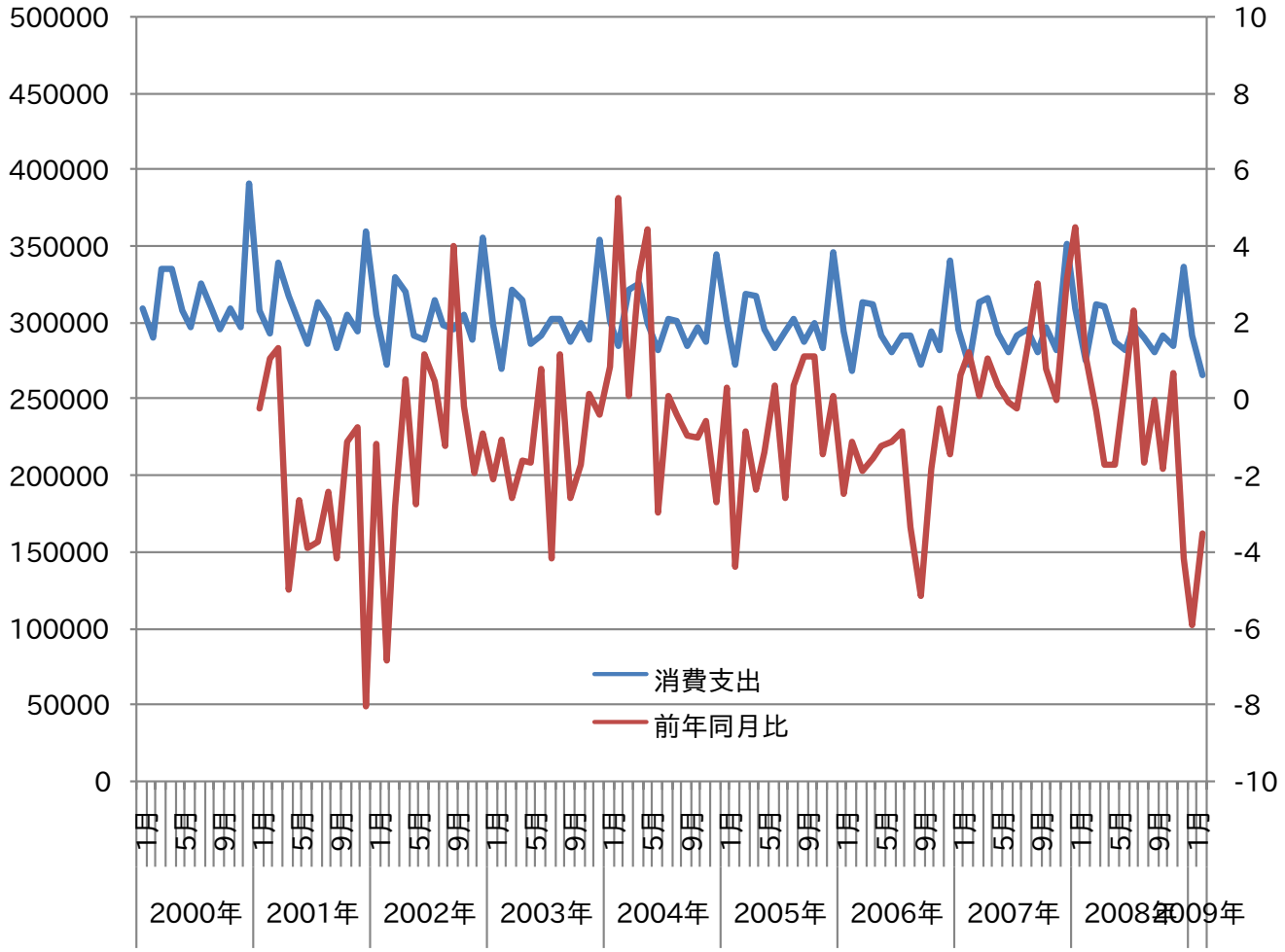
- 「地域」の定義に注意
  - どのようにそのデータが作られたのか
  - 越境の可能性は？　分析の目的に即しているか？
    - 昼間人口と夜間人口，本社所在地など
  - 合併や統合が起きていないか？



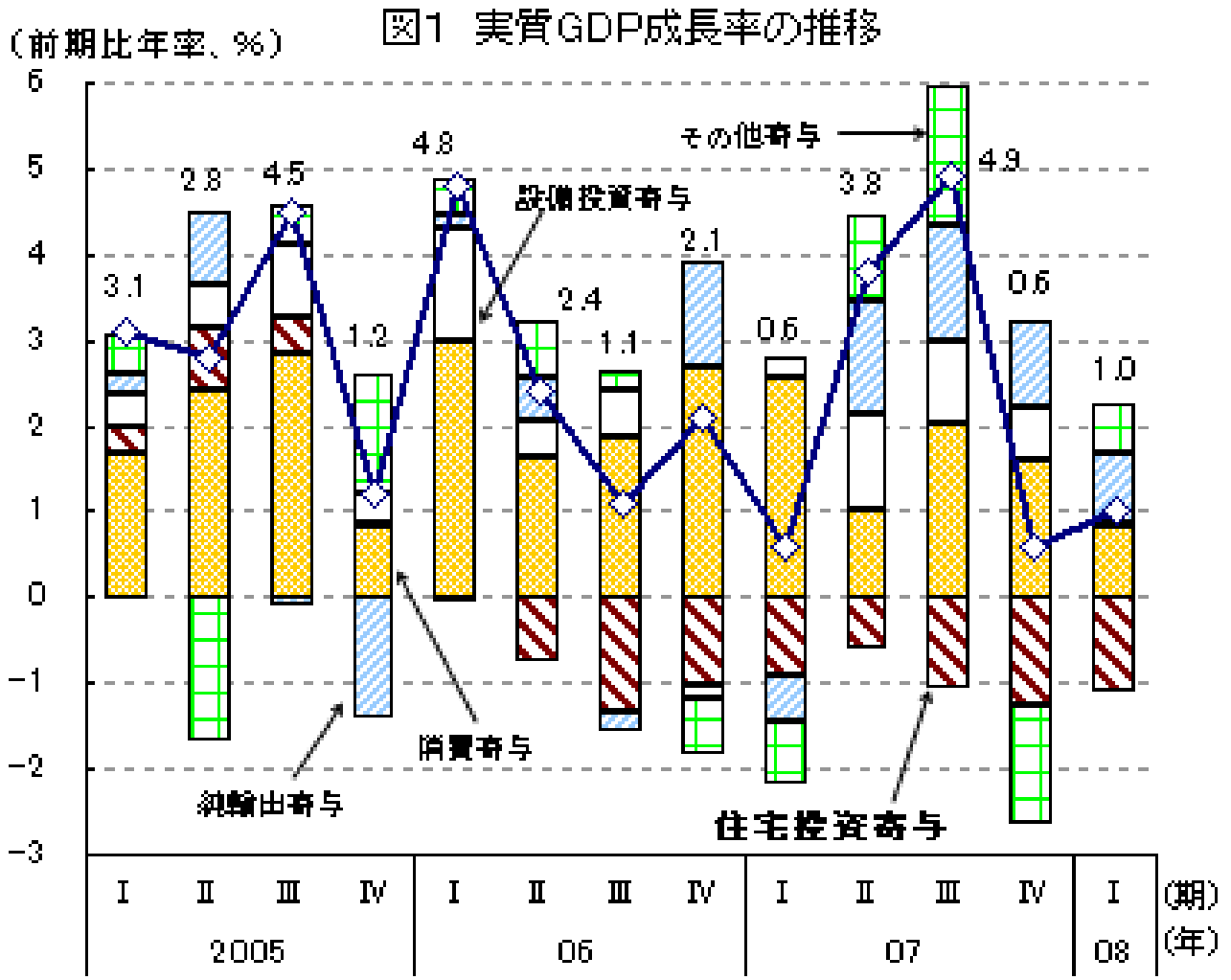
基本的な変形:時系列

- 変化率
  - 時系列のデータはしばしばトレンドや季節性をもつ
  - それらが共通の要因として重要でないときには変形が必要
- いくつかの変化率
  - 成長率 = $x_t / x_{t-1} - 1$
  - (四半期の場合)前年同期比= $x_t / x_{t-4} - 1$
  - (月次の場合)前年同月比= $x_t / x_{t-12} - 1$
- リードとラグ
- 寄与度分解

基本的な変形:季節性(家計消費)



寄与度分解（内閣府「今週の指標」887，2008年7月22日）

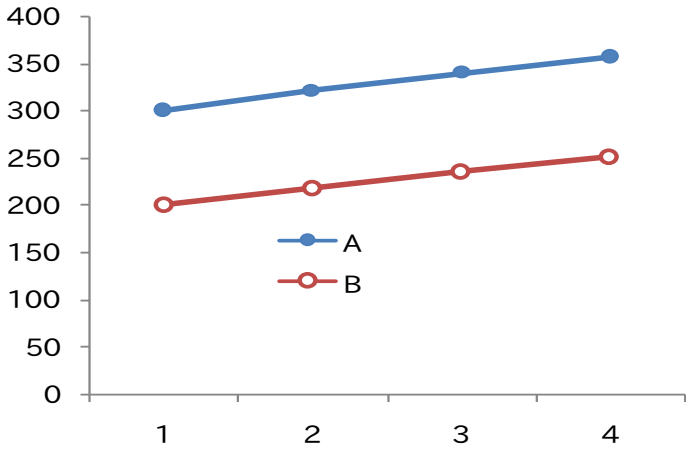


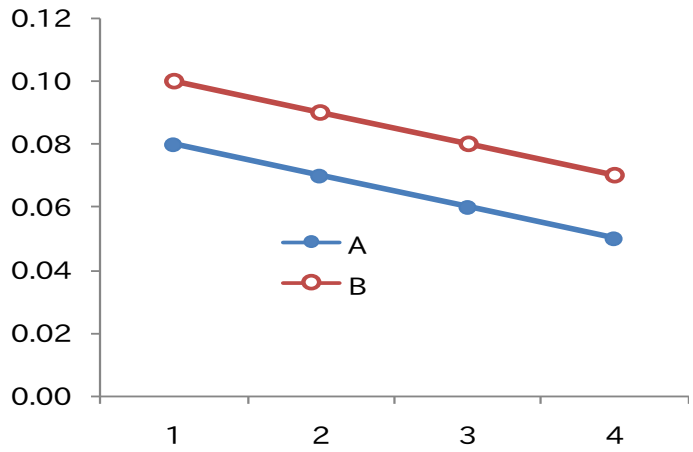
基本的な変形:時系列

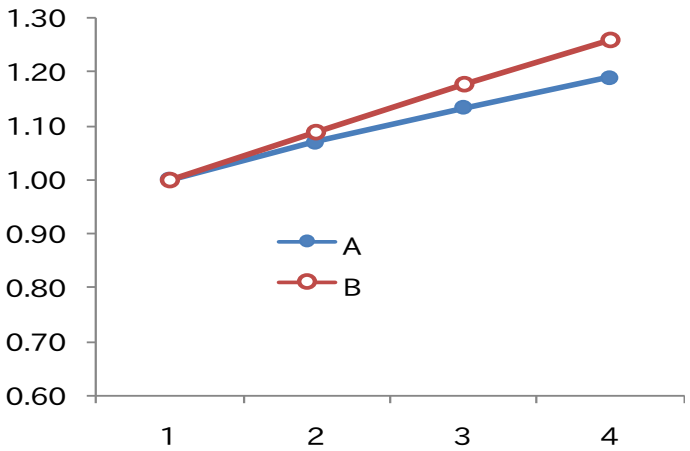
- 指数
  - 時系列などのデータを加工して,比較しやすくしたもの
  - いくつかの系列の動きをまとめて示すために計算される
- 簡単な指数
  - 「ある時点を100として」:出発点が異なるものの比較 = x_t / x_1
  - 例:株式市場間の動きの比較などの地域間指数
- 複雑な指数
  - 例:物価指数,鉱工業生産指数
  - 作成者が定義してしまう指数:株価指数など
  - 質的変数を変形した指数:日銀DIなど
- 変化率を見たいのか,水準を見たいのかに注意.
  - 定義を調べて,適切な指数を選ぼう.

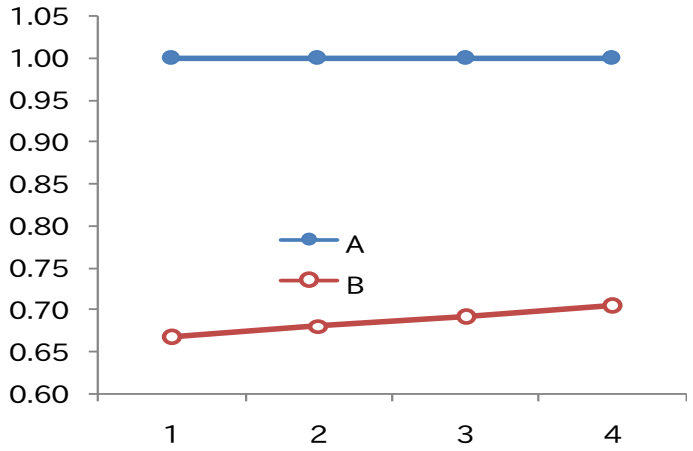
指数と変化率

- 何がわかるか?／何を言いたいのか?









時系列の注意点：ストックとフロー

- ストックとフロー
  - ストック：ある時点での量を示す（年度末時点など）
  - フロー：ある期間内での量を示す（年度など）
  - フローの積み重ねがストックになる
  - 例：政府の借り入れ：公債残高はストック，公債収入はフロー
- 区別する方法
  - もとの資料に当たって，変数の定義を調べる
  - ストックなら「～時点」，フローなら「～中」とあるはず
- ストックとフローの散布図も有用なことも
  - 発生率 = ある期間の発生率／期首時点での数
  - 分母が何かに注意
- ストック／フロー　で，平均時間を求めることもある