

基礎統計学

記述統計(2)
さまざまな指標 & 2変数データの整理

京都大学経済学研究科
岩城秀樹

経済統計(1)

- 社会統計学の立場
 - 社会科学に基礎をおいた統計学
 - 社会を対象 社会を数量的に認識、把握、分析、推測するための方法論
- 経済統計
 - 経済の諸現象の量的な側面を反映するデータの集まり
 - 経済状況、社会状況を調べる、知るための指標
 - 人口統計、家計調査、生産統計、物価指数、金融統計、国民経済計算(SNA)など
 - 国は、様々な調査を実施し経済統計資料を有している。

経済統計(2)

- 経済統計の分析には、その目的や特性に合わせた収集、集計、指標が必用
- 調査統計(指定統計)
 - 法律(統計法)に基づき、被調査者の調査への協力が法的に義務づけられている。
 - 調査方法によって、全数調査と標本調査
- 業務統計
 - 政府の業務統計
 - 人口動態統計、貿易統計、犯罪統計、金融統計など
 - 総務省統計局 <http://www.stat.go.jp/>
 - 統計データ・ポータル <http://portal.stat.go.jp/PubStat/top.html>

統計調査の手順

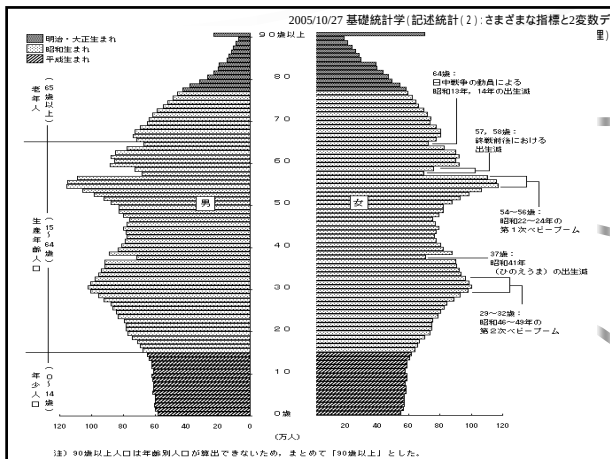
- 統計調査を行うために 誰、いつ、どこで、何を、如何に
 - 主体:調査主体、調査目的
 - 調査時期と調査時点:
 - 場所:調査の対象地域の限定
 - 対象:何を調査対象とするか、範囲を明確に規定する。
 - 調査単位:個人、世帯、事業所
 - 調査方法:全数調査か、標本調査か
 - 全数調査、有意抽出法(調査者が適当に標本を選ぶ)、無作為抽出法(ランダムに抽出)

統計指標

- (一般的な)代表値だけでは不十分、個々の調査データにあわせた指標
 - 人口統計:人口、出生率、死亡率、合計特殊出生率、コーホート変化率など
 - 家計調査:所得、貯蓄率、不平等指標(ジニ係数)、エンゲル係数など
 - 生産統計:GNP、GDP、景気指標(業況DI(Diffusion Indexes)、景気動向指標)など
 - 物価指数:ラスパイレス指標、GDPデフレーター、海外物価指標、購買力平価など
 - 金融統計:金利、流動性、M2+CDなど

人口統計(1)

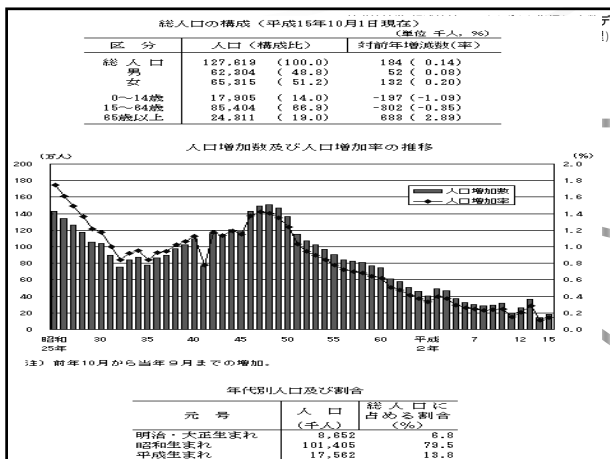
- 人口静態統計:人口の規模、構成
 - 国勢調査(人口センサス)、5年に1度、10月1日
 - 年齢グループ
 - 年少人口(0~14歳)、生産年齢人口(15~64歳)、老年人口(65歳以上)
 - 年少人口指数 = $\frac{\text{年少人口}}{\text{生産年齢人口}} \times 100$
 - 老年人口指数 = $\frac{\text{老年人口}}{\text{生産年齢人口}} \times 100$
 - 従属人口指数 = $\frac{(\text{年少人口} + \text{老年人口})}{\text{生産年齢人口}} \times 100$
 - 老年化指数 = $\frac{\text{老年人口}}{\text{年少人口}} \times 100$
 - 老年人口比率 = $\frac{\text{老年人口}}{\text{総人口}} \times 100$



2005/10/27 基礎統計学(記述統計(2)):さまざまな指標と2変数データの整理

人口統計(2)

- 人口動態統計: 出生・死亡などの人口再生産
 - 出生率 = 年間出生数/10月1日総人口 × 1000
 - 死亡率 = 年間死亡数/10月1日総人口 × 1000
 - 自然増加率
 - = (年間出生数 - 年間死亡数)/10月1日総人口
 - 使用単位: 千分率、パーミル(Per mille)
 - 合計特殊出生率(Total Fertility Rate)
 - 1人の女性が生涯に生む子供の数
 - $TFR = \sum \{B(x)/F(x)\}$
 - $B(x)$: 女子1人あたり年齢別出生数(=出生率)
 - $F(x)$: 年齢別女子人口



2005/10/27 基礎統計学(記述統計(2)):さまざまな指標と2変数データの整理

人口統計(3)

- 人口移動統計: 人口の移動
 - 一定期間、一定地域における人口増加は自然増加と社会増加からなる。
 - 自然増加 = 出生数(B) - 死亡数(D)
 - 社会増加 = 転入数(I) - 転出数(O)
- 人口の推計: コーホート変動
 - コーホート: Cohort, 古代ローマ軍歩兵隊(300-600人)
 - 人口の同時発生集団: 同一期間に出生した同年齢の出生集団を出生コーホート変動という。

2005/10/27 基礎統計学(記述統計(2)):さまざまな指標と2変数データの整理

家計調査(1)

- 家計調査
 - 国民生活の実態を家計収支の面から明らかにすることが目的
 - 世帯毎の収入、支出、所得、貯蓄、負債など
- 収入総額 = 実収入 + 実収入以外の収入 + 繰入金
- 支出総額 = 実支出 + 実支出以外の支出 + 繰越金
- 可処分所得 = 実収入 - 非消費支出

2005/10/27 基礎統計学(記述統計(2)):さまざまな指標と2変数データの整理

家計調査(2)

- 貯蓄 = 可処分所得 - 消費支出
 - = (実収入 - 非消費支出) - 消費支出
 - = 実収入 - 実支出
 - = (実支出以外の支出 - 実収入以外の収入) + (繰越金 - 繰入金)
 - 最後の式は収支がバランスしているという事実から導かれる。
- 可処分所得に占める消費支出の割合を平均消費性向、貯蓄の割合を貯蓄率という。
- 貯蓄率 = 黒字率 = 貯蓄/可処分所得
 - = 1 - 平均消費性向

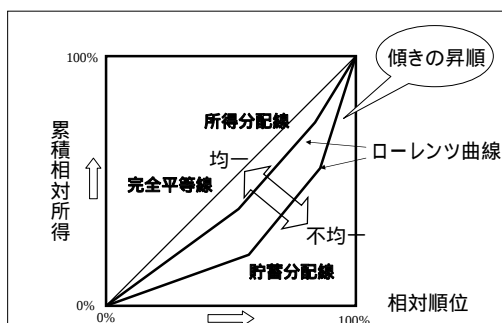
家計調査(3)

- 世帯の富の分配、集中を調べる指標
 - ダンカンの(人口)集中係数
 - $\sum | \text{地域面積の割合} - \text{人口の割合} | / 2$
ある地域の構成割合の差の絶対値の和を2で割る
 - ローレンツ曲線
 - ジニ係数
- 相対順位: 総人数の中での順位の割合
- 相対所得: 全体所得における占める個人所得の割合
- 累積相対所得: ある順位までの相対所得の和

ローレンツ(Lorenz) 曲線(1)

- アメリカの官庁統計家であったロ・レンツ(M.O.Lorenz)が考案
- 相対順位と相対所得などのシェアを用いて経済格差(不平等度)を表現するグラフ
 - 横軸に相対順位、縦軸に累積相対所得(累積シェア)をとることによって、グラフは弓形に描かれるが、その弓形の大きさから経済格差(不平等度)の程度が分る
 - 所得格差がなければ、相対所得(所得/全員の所得の和)と相対順位(順位/人数)は一致する 完全平等線(45度線)

ローレンツ(Lorenz) 曲線(2)

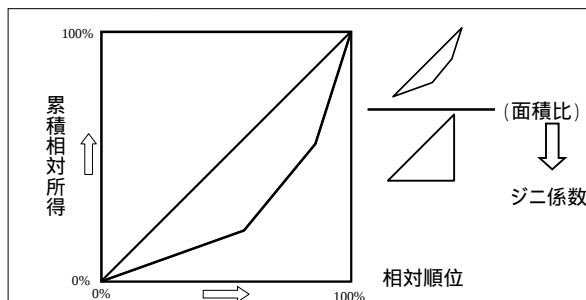


ジニ係数(Gini coefficient)(1)

- イタリアの統計学者コッラド・ジニ(Corrado Gini)が考案
- 主に社会における所得分配の不平等さを測る指標。ローレンツ曲線を描くことで求められる。
 - 係数の範囲は0から1で、係数が大きくなるほど、不平等さが高い。
 - 0のときには完全な「平等」: 全員が同じ所得
 - 1のときには完全な「不平等」: 1人が全所得を独占
 - 所得分配の不平等さ以外にも富の偏在性やエネルギー消費における不平等さなどに応用

ジニ係数(2)

- ジニ係数 = $1 - \frac{\text{ローレンツ曲線下の多角形の面積}}{\text{三角形の面積}}$



ジニ係数(3)

- ジニ係数 = $1 - \frac{\text{ローレンツ曲線下の多角形の面積}}{\text{三角形の面積}}$
- = $2 \times \{ \text{相対順位と相対所得の積の総和} - (n+1)/n \}$
- 不平等さを客観的に分析、比較する際の代表的な指標のひとつ
- しかし、同じジニ係数で示される状態であっても、ローレンツ曲線の元の形が著しく違えば、実感として感じる不平等さがまったく変わってくる可能性もあるため、注意が必要である。

家計調査(4)

■富の分布

- 所得での比較では単年度のみ 貯蓄を考慮
- グループの相対貯蓄額

$$= \frac{\text{そのグループの相対度数} \times \text{平均貯蓄額の積}}{\sum \text{そのグループの相対度数} \times \text{平均貯蓄額}}$$
- 相対順位 = そのグループ以下の相対度数の和/100
- 貯蓄を配慮しても不十分 資産の分配
- 固定資産の評価が困難
- ローンなどの負債を考慮することが困難
- 真の富の分布を調べることは非常に難しい

生産統計(1)

■GNP(国民総生産:Gross National Product)

- 国全体の経済の大きさを測る指標の一つ、一定期間内に生産した物を市場価格で評価した値から、中間生産物を差し引いて求める値
- 原則として、国内または海外を問わず、日本国民により生産された生産額の総計

■GDP(国内総生産:Gross Domestic Product)

- 国内総生産 = 付加価値の合計 = 消費 + 投資
- 国全体の経済活動を総合的に表す指標の一つ、付加価値の合計で、消費金額に一致しており、構成要素中の個人消費は、国内総生産

生産統計(2)

■名目GDP = 「現行価格表示のGDP」

- $GDP_t = \sum P_{ti} Q_{ti}$ = {その年(t)の財iの値段 × その年に取引された財iの数量}

■実質GDP = 「不変価格表示のGDP」

- $GDP_0 = \sum P_{0i} Q_{ti}$ = {基準年の財iの値段 × その年に取引された財iの数量}
- ある基準時点の価格体系を設定し、その基準年の価格で評価したGDP
- $GDP_0 = \sum (P_{0i}/P_{ti}) S_{ti}$
 - 基準年価格で評価したt年の実質販売額の和
 - 価格上昇率 P_{ti}/P_{0i} 、t年における名目販売額 S_{ti}

物価指標(1)

■価格

- 現行価格(current price):日々の市場で取引される値段
- 不変価格(constant price):基準時点の価格(現在は2000年)に変換した価格

■物価指数

- 商品の平均的な値上がり率
- 加重平均の応用例 支出割合の比率に応じて重みをつけた値上がり率
- ラスパイレ方式(Laspeyres,L式)
- パーシェ方式(Paasch,P式)

ラスパイレ(Laspeyres)指数

- 基準年の支出割合に応じて個々の財の価格指標の重みをつけた値上がり率

$$L_t = \sum \left\{ w_0 \times \frac{p_t}{p_0} \right\} = \sum \left\{ \frac{p_0 q_0}{\sum p_0 q_0} \times \frac{p_t}{p_0} \right\} = \frac{\sum p_t q_0}{\sum p_0 q_0}$$

$\frac{p_t}{p_0}$: 当該財の価格指数

$w_0 = \frac{p_0 q_0}{\sum p_0 q_0}$: 当該財の基準時ウェイト

購入量を基準年に固定し、この組み合わせを購入するのに必要な支出額を比較

物価指標(2)

■消費者物価指数(CPI)

- 全国の消費者世帯が購入する各種の商品・サービスの価格動向を示すもの。

■卸売物価指数(WPI)

- 企業間で取引される物的商品(サービスは除く)の価格動向を示すもの。
- 採用品目とウェイトを基準時点で固定するラスパイレ指数

パーシェ(Paasch)指数(1)

- 当該年の支出割合に応じて個々の財の基準価格指標の重みをつけた値上がり率

$$P_t = \sum \left\{ \frac{p_0 q_t}{\sum p_0 q_t} \times \frac{p_t}{p_0} \right\} = \frac{\sum p_t q_t}{\sum p_0 q_t}$$

$$= \frac{1}{\sum \frac{p_0 q_t}{p_t q_t}} = \frac{1}{\sum \frac{p_0}{p_t} \frac{p_t q_t}{q_t}} = \frac{1}{\sum \left\{ \frac{p_0}{p_t} \times \frac{p_t q_t}{q_t} \right\}}$$

パーシェ指数(2)

$$P_t = \frac{1}{\sum \left\{ \frac{p_0}{p_t} \times \frac{p_t q_t}{\sum p_t q_t} \right\}} = \frac{1}{\sum \left\{ w_t \times \frac{p_0}{p_t} \right\}}$$

$\frac{p_0}{p_t}$: 当該財の価格指数の逆数

$w_t = \frac{p_t q_t}{\sum p_t q_t}$: 当該財の比較時ウェイト

- 個々の財・サービスの価格指数を比較時のウェイトで調和平均したもの

パーシェチェック

- 定期的にP式指数を計算してL式指数との差(1 - L/P)を調べ、現在使用しているL式指数の信頼性をチェックする

結果に大きな差があれば、基準時点を変更する必要が生じる。

- 5年毎のパーシェ指数を求めてラスパイルズ指数との差を調べる(経済企画庁)

- ラスパイルズ指数(L式指数)はパーシェ指数(P式指数)よりも時間が経つにつれて、大きな上昇を記述する傾向がある。

- 基準年度で支出割合を決めるため、価格が高くなっても消費量を減らすという消費の変化が反映されない

物価指標(3)

- 海外物価指標

- 為替レートを用いて商品セットの購入するのに必要な金額の比を求める

$$P_{AJ} = \sum P_{Ai} Q_i / \sum P_{Ji} Q_i$$

- 日本の物価(J国)を基準としたA国の物価指標
- P_{Ai} は、為替レートを用いて円評価したもの

- 購買力平価(PPP: Purchasing Power Parity)

- 商品セットを購入するのに必要な金額の比を求めて為替の価値を求める

$$PPP = \sum P_{Ji} Q_i / \sum P_{Ai} Q_i$$

GDPデフレーター

- 経済全体の価格指標

- GDPの物価指標(パーシェ指数)

- GDPデフレーター

$$= \text{名目GDP} / \text{実質GDP} = \sum P_{ti} Q_{ti} / \sum P_{0i} Q_{ti}$$

- 実質GDP₀ = $\sum (P_{0i} / P_{ti}) S_{ti}$

- 基準年価格で評価したt年の実質販売額の和

- 価格上昇率 P_{ti} / P_{0i}

- t年における名目販売額 S_{ti}

- 名目GDP_t = $C_t + I_t + G_t + (X_t - M_t)$

C:消費、I:投資、G:政府支出、X:輸出、M:輸入

金融統計

- マネーサプライ(通貨供給量: Money Supply)

- 基本的に通貨保有主体(一般法人、個人、地方公共団体など)が保有する通貨量の残高 $M2 + CD$

- M2(現金通貨、預金通貨、準通貨)にCD(譲渡性預金)を加えたもので通貨発行主体(日銀、国内銀行など)が発行し、通貨保有主体が保有する現金・預金の集計

- マーシャルのK

- マネーサプライ($M2 + CD$)/名目GDP

- 経済活動(GDP)にとって、マネーサプライが適正水準にあるかどうかを判断するための指標

- 貨幣流通速度: 名目GDP/($M2 + CD$)

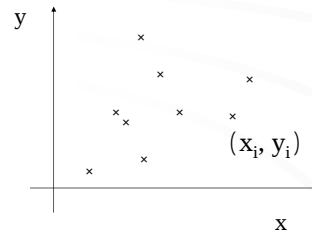
- 経済取引を媒介するために、一定期間に流通しなければならない回転数

2変数データ

- 一般に、対象となる標本は複数の特性を有している
 - 例: 人間 → 身長、体重、性別、など
 - 例: 世帯 → 世帯数、貯蓄額、所得、など
- 2個(複数個)の特性(データ)を同時に扱う
2変数データ
 - 特性 x と特性 y 、観測個数を n とすると、
 $(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)$
 と表現する
 - 数学的には2変数以上を表すのでベクトルとなる

散布図

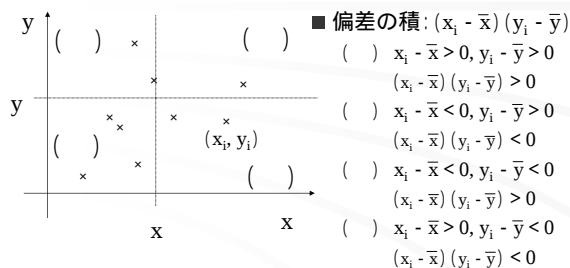
- 2変数データをそれぞれ x 軸、 y 軸に対応させて点をプロットする。
 $(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)$



x と y の間に何らかの関係がある。ただし、どちらが原因であり、どちらかが結果といえない:
因果関係がない

相関(Correlation)関係

相関関係(1)



$(x_i - \bar{x})(y_i - \bar{y}) > 0$ ならば、に点が存在 右上がり(左下がり)
 $(x_i - \bar{x})(y_i - \bar{y}) < 0$ ならば、に点が存在 右下がり(左上がり)

相関関係(2)

- 積和: $\sum (x_i - \bar{x})(y_i - \bar{y})$
- 正の相関
 - 、に点が存在 右上がり(左下がり)
 $(x_i - \bar{x})(y_i - \bar{y}) > 0$ の数が多い $\sum (x_i - \bar{x})(y_i - \bar{y}) > 0$
- 負の相関
 - 、に点が存在 右下がり(左上がり)
 $(x_i - \bar{x})(y_i - \bar{y}) < 0$ の数が多い $\sum (x_i - \bar{x})(y_i - \bar{y}) < 0$



相関関係(3)

- 相関の強さを表す指標
 データ: $(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)$ とすると
- 共分散 σ_{xy} = 積和/ n

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$= \frac{\sum x_i y_i}{n} - \bar{x} \bar{y}$$
- 標本共分散 s_{xy} = 積和/ $(n - 1)$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

$$= \frac{\{\sum x_i y_i\} - n \bar{x} \bar{y}}{(n - 1)}$$

$x=y$ のとき、共分散 $\sigma_{xy} = \sigma_x^2$
 標本共分散 $s_{xy} = s_x^2$

相関関係(4)

- 相関の強さを表す指標
 相関係数(ピアソンの積率相関係数)
 (Productmoment Correlation Coefficient)
- 標本相関係数 = $\frac{\text{標本共分散}}{\sqrt{x \text{の標本分散} \times y \text{の標本分散}}}$

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}) / (n - 1)}{\sqrt{\sum (x_i - \bar{x})^2 / (n - 1)} \sqrt{\sum (y_i - \bar{y})^2 / (n - 1)}}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

■ -1 r_{xy} 1

相関係数 (5)

■相関係数の値と関係

- $r_{xy} = 1$ 完全な正の相関
 - 右上がりの直線
- $0.7 < r_{xy} < 1.0$ 強い正の相関
- $0.3 < r_{xy} < 0.7$ 弱い正の相関
- $-0.3 < r_{xy} < 0.3$ 無相関
- $-0.7 < r_{xy} < -0.3$ 弱い正の相関
- $-1 < r_{xy} < -0.7$ 強い正の相関
- $r_{xy} = -1$ 完全な負の相関
 - 右下がりの直線

相関係数の注意 (1)

■相関係数と因果関係

- 「相関」と「因果」は異なる概念 相関が高いからと言っても因果関係があるわけではない
- 例えば、身長と体重は、一般に高い相関係数がある。しかし、「身長が高い(原因)から体重が重い(結果)」を示せるわけではない

■外れ値の扱い

- まとまりから大きく外れた値は、本来のまとまりを持つ相関係数と異なる結果を招くことがある
- $(1,1) \dots (9,9)$, $(10,10)$ の相関係数は1であるが、1つだけ大きな外れ値があった場合、 $(1,1) \dots (9,9)$, $(10,100)$ になるだけで相関係数は0.59と大きく変化

相関係数の注意 (2)

■層化の問題

- グループ毎に異なる関係がある集団が混ざってしまうと相関がなくなるように見える
- $(1,1) \dots (10,10)$ と $(1,10) \dots (10,1)$ という2系統のグループの2変数データが混在していると、グループ毎には強い関係が見られるが、混在すると結果として相関がないように見えてしまう。

■線形とは異なる関係

- 相関係数は、あくまでも線形な関係のみを対象
- 直線上の関係があるかどうかのみをチェック
- データが円周上 ($x^2 + y^2 = 10$) に配置されているとすると、相関はないが、何ら関係がないとは言えない。

2変数同時度数分布

■2変数同時度数分布/分割表

- 縦軸と横軸にそれぞれの特性をとって、データを大きさによっていくつかの組に分け、各級に入るデータ数でマス目の表をつくったもの
- 分割表から、変数毎の和を求めると各変数の度数分布を導ける 周辺度数分布

性別 \ 年齢	0 - 15	15 - 60	60 -	性別の度数分布
男性	100	230	80	410
女性	80	300	50	430
年齢の周辺度数分布	180	530	130	840

標本偏相関係数

■見せかけ相関/擬似相関

- 直接関係がないにもかかわらず、間接的な変数が関わることで相関があるように見える

■標本偏相関

- 「他の条件を一定とした場合の2変数の相関係数の強さ」を表わす尺度
- x , y , z の3変数で z 変数の影響を除去した x と y の標本偏相関係数

$$r_{xy|z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

順位相関係数

■2つの変数 x , y が順位を表す場合

- 変数データは自然数
- $\bar{x} = \bar{y}$ (同一順位はないものとする)

■データ数が n 個とすると、

1から n までの自然数の分散なので

$$s_x^2 = s_y^2 = (n^2 - 1)/12 \text{ であることを利用すると}$$

$$r_{xy} = 1 - \frac{6 \sum (x_i - y_i)^2}{n(n^2 - 1)}$$

となる

回帰関係 (1)

■回帰分析 (Regression Analysis)

- 原因 (Cause) と結果 (Result) の関係を、法則性としてとらえる統計的方法

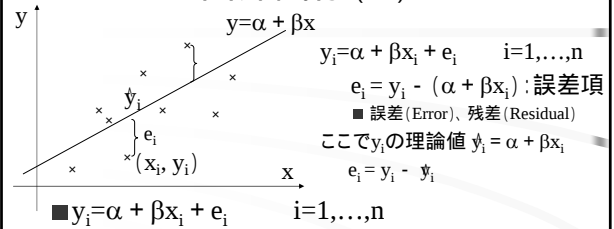
■線形回帰 (Linear Regression)

- 2つの変数が線形の関係

$$Y = \alpha + \beta X$$

- Y: 被説明変数 (従属変数) ← 結果変数: 縦軸
- X: 説明変数 (独立変数) ← 説明変数: 横軸
- 回帰とは、結果 (Y) を原因 (X) に説明させる
 - 線形以外の関係場合、非線形回帰と呼ぶ
 - 説明変数が複数ある場合、重回帰と呼ぶ

回帰関係 (2)



$$y_i = \alpha + \beta x_i + e_i \quad i=1, \dots, n$$

■残差の条件 (性質)

- $E(e_i) = 0, \text{Var}(e_i) = \text{Var}(y_i) = \sigma^2$ (分布は独立)
 - より詳細な条件や性質があるが省略。計量経済学で必須
- 実測値 $(x_1, y_1) \dots (x_n, y_n)$ から、どうやって合理的に α, β を求めるか 最小二乗法

最小二乗法 (1)

■最小二乗法 (Least Square Method)

- 残差二乗和: RSS (Residual Sum of Squares) が最小となるような回帰線を推計する方法
 - 残差の二乗和を最小にするように α, β を決めるのが合理的。他にも最尤法なども存在
 - 一般の最小二乗法は、OLS (Ordinary Least Square Method) と略される

$$y_i = \alpha + \beta x_i + e_i \quad (i=1, \dots, n) \text{ としたとき}$$

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (\alpha + \beta x_i))^2 \rightarrow \text{Min}$$

を求める

$$\frac{\partial d}{\partial \alpha} \sum \{y_i - (\alpha + \beta x_i)\}^2 = 0 \quad \frac{\partial d}{\partial \beta} \sum \{y_i - (\alpha + \beta x_i)\}^2 = 0 \text{ を解く}$$