経済統計分析 6
確率の基礎

第1回宿題について

- 答えが合わないのですけど?
  - ひょっとして違うデータを使っているかしら?
- 成長率（変化率）の定義
  - $(x_t - x_{t\text{-}1})/x_{t\text{-}1}$ が一般的. %表示にしてもよい.
- 記述統計の計算
  - 組込み関数を使うとよい. count, average, sumなど.
  - 標準偏差については,「今回は」STDEVPを使います.
  - もちろん, 差をとって2乗して足して割ってもよいです.
- ローレンツカーブ
  - 各国を「階級値」と考えてもらいたかったところです.
  - 散布図を使って描きましょう.

今日のおはなし.

- 統計的推測 statistical inference へ向けての準備
  - 確率論の基礎用語
  - 確率分布, とくに正規分布
  - 条件付き分布
  - 標本分布

- 今日のタネ
  - 吉田耕作. 2006. 直感的統計学. 日経BP.
  - 中村隆英ほか. 1984. 統計入門. 東大出版会.

なにができるようになりたいか

- ある変数が他の変数に与える効果の大きさの数量化
  - 例:「統治状況」は一国経済の成長率に, 平均的には, どれほど影響するのか?
  - 例:「統治状況」がいいところと悪いところでは経済成長率に差があるのか?
- でも, 社会経済事象にはさまざまな要因が影響する
  - いくつかの事象は捨象せざるをえない
  - すべてのデータを集めることは不可能
- だから, 観察されるデータに誤差や散らばりはつきもの
  - なんらかの意味での「でたらめさ randomness」がつきまとう
  - 例:経済成長率は「統治状況」だけに影響されるわけではない
  - 例:ある年の経済成長率のデータは特殊要因に左右されるかも
- 「でたらめさ」を扱う手法が必要　→確率論.

確率論の考え方

- 確率論
  - 「不確かさ」や「リスク」を扱うための数学的手法
  - 将来起きうることを列挙し, それぞれの「起きやすさ」を数値で表現
  - 「起きやすさ」って?
- 先験的(理論的)確率
  - 起きうることがいくつかあるとき, どれがとくに起きやすいと考える理由がないとき, それらの「起きやすさ」はすべて等しいと考えよう
- 経験的(実験的)確率
  - これまでの経験や実験から, それぞれのできごとの起きる相対頻度が分かっており, 一定の値に収束すると思われるとき, その収束先を「起きやすさ」と考えよう
- 主観確率
  - 確信の度合い, 信念などによって「起きやすさ」を主観的に割り振る
  - 意思決定の前段階として位置づけられることが多い

事象, 根源事象

- 「起きうること」を全て挙げたとき,
  - 「起きうること」を一般に事象 event と呼ぶ
  - 根源事象 : 相互に排他的で, それらの組合せによって他の「起きうること」を表現できるような事象
  - 標本空間 (Ω) : 根源事象全てから成る集合.
  - 空事象 (ϕ) : なにも起きないこと
- 例 : サイコロ投げ
  - 事象 : 「ピンの目が出る」「4以上の目が出る」「偶数の目が出る」……
  - 根源事象 : 「1」「2」「3」「4」「5」「6」
  - 標本空間 : 「1か, 2か, 3か, 4か, 5か, 6の目が出る」
  - 空事象 : なにも起きない

確率が満たすべき条件

- 確率が満たすべき条件
  - 任意の事象Aに対して，$0 \leq \Pr(A) \leq 1$
  - 標本空間と空事象に対して，$\Pr(\Omega) = 1$, $\Pr(\phi) = 0$.
  - 相互に排他的な事象A_1, A_2に対して，$\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2)$
- 例：サイコロ投げの先験的確率
  - 根源事象の確率はそれぞれ1/6：標本空間の確率が1だから．
  - 事象Aを「4以上の目が出る」
  - $\Pr(A_1) = \Pr(4) + \Pr(5) + \Pr(6) = 1/6 + 1/6 + 1/6 = 1/2$
- 確率の加法公式
  - $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2)$
  - ベン図を描こう

確率分布

- 確率変数
  - 「でたらめ」の実現に応じてさまざまな値を取る変数
  - 確率変数そのものを大文字, 実現値を小文字で書く習慣
  - 例 : さいころの出る目を X で表し, $\Pr(X=1) = 1/6$
- 確率分布
  - 根源事象と, 対応する確率の一覧
  - 例 : サイコロ投げの確率分布

目	1	2	3	4	5	6
確率	1/6	1/6	1/6	1/6	1/6	1/6

  - 例 : 2個のサイコロ投げの確率分布

目	1, 1	1, 2	1, 3	1, 4	1, 5	1, 6	…	6, 5	6, 6
確率	1/36	1/36	1/36	1/36	1/36	1/36	…	1/36	1/36

  - 連続変数のばあい, このような確率分布は考えにくい

累積確率分布, 確率密度

- 累積確率分布c.d.f. (cumulative distribution function)
  - 確率変数がある値より小さな値を取る確率 $f_X(x)$ = Pr($X \le x$)
  - 例: サイコロ投げの累積分布関数

目	1	2	3	4	5	6
分布	1/6	2/6	3/6	4/6	5/6	1

- 確率密度 p.d.f. (probability density function)
  - 連続関数のときだけ
  - 累積分布関数の微分値　→全区間について積分すると1
  - 確率分布の棒グラフの高級な(?)やつだが, 1以上の値も取りうる

同時分布, 周辺分布

- 同時分布 joint distribution
  - 2つ以上の確率変数があるとき, それらの実現値の組合せにたいする確率の一覧
  - 例: 2枚前のスライド
  - 例: 天気と通勤時間（Stock and Watson, Table 2.2.）

	雨（X=0）	晴れ（X=1）	
遅れる（Y=0）	0.15	0.07	
遅れない（Y=1）	0.15	0.63	

  - 実現値の組合せが「事象」となるので, 同時確率の和が1
- 周辺分布 marginal distribution
  - 同時分布が与えられたときに, 1つの変数だけに着目してえられる確率分布

条件付き分布

- 条件付き確率, 条件付き分布 conditional distribution
  - 2つ以上の確率変数があるとき, ある確率変数の実現値を所与としたときの（ある確率変数で条件付けしたときの）他の変数の確率分布
  - 例：天気で条件付けたときの通勤時間の条件付き分布

	雨（X=0）	晴れ（X=1）
遅れる（Y=0）		
遅れない（Y=1）		
	1	1

  - 一般的には, 条件付き分布は, 条件付けた変数の関数となる
- 記法
  - 同時確率：$\Pr(X = x, Y = y)$
  - 条件付き確率：$\Pr(X = x | Y = y) = \Pr(X = x, Y = y) / \Pr(Y = y)$

ベイズの定理

- 条件付き確率と同時確率の関係

  $\Pr(X = x, Y = y) = \Pr(Y = y) \Pr(X = x \mid Y = y)$

  $\Pr(X = x, Y = y) = \Pr(X = x) \Pr(Y = y \mid X = x)$

- 左辺は同じものだから,

  $\Pr(Y = y) \Pr(X = x \mid Y = y) = \Pr(X = x) \Pr(Y = y \mid X = x)$

- 両辺を割ってみると, ベイズの定理をえる

  - 右辺と左辺で, 条件付けされている変数が入れ替わっていることに注意!
  - $\Pr(Y = y) = \sum \Pr(Y = y \mid X = x_i) \Pr(X = x_i)$ という関係を使って変形できる

ベイズの定理の応用例

- ベイズ流の情報のアップデートの例
  - 右辺に入っている Pr(X) が事前確率
  - 左辺でもとまる Pr($X = x \mid Y = y$) が事後確率
- ○×式の試験結果から, 理解しているかどうかを推測
  - X : 答えが分かっているかどうか. 分かっていれば1, いなければ0
  - Y : 試験に正答すれば1, 間違えれば0
  - 仮定 : Pr($Y = 1 \mid X = 1$) = 1, Pr($Y = 0 \mid X = 1$) = 0
  - 仮定 : Pr($Y = 1 \mid X = 0$) = 1/2, Pr($Y = 0 \mid X = 0$) = 1/2
  - 右辺はPr(X) の関数として表現できる

  - たとえば, Pr($X = 1$) = 1/2 のとき, $Y = 1$ なら, Pr($X = 1 \mid Y = 1$) = 2/3
  - 「正答した」という情報から「分かっている」確率が上方修正された

独立

- 独立 independent
  - 2つの変数が独立であるとは, すべての起きうる値に対して, 条件付き分布が周辺分布に等しいことをいう.
  - このとき, 条件付き分布の定義より, 同時分布は周辺分布の積
  - 例: 2つのサイコロ投げ
  - 2つの確率変数は相関を持たない
  - 片方の確率変数の実現値の情報が分かったとしても, もうひとつの確率変数の確率分布について新たな情報とならない

確率分布の特性値

- 確率分布がすでに分かっているとする
- 確率分布の特性値
  - 確率分布の状況を特徴付けるような数値
  - 確率分布の記述統計量といってもよい
  - よく使うのは（条件付き）平均と分散

- 確率分布の状況が分かっていないとき，
  - 特性値の値が，「統計的推測（推定）」のターゲットとなる
  - 一般に，手元にあるデータから確率分布を完全に復元するのは不可能
  - じっさい，「統計的推測」とは，分かっていない特性値を推測することといってよい
  - 例：日本の平均賃金率．学歴別の平均賃金率．

平均, 分散

- 平均 mean, average
  - 確率を重みとする実現値の加重平均

$$\mu = E[Y] = \sum_{i=1}^{k} y_i \Pr(Y = y_i)$$

  - 連続変数のばあいも, 加重平均みたいなもの(積分値)
- 分散 variance
  - 確率分布の「広がり」「散らばり」を表す
  - 各実現値から平均を引いたものの2乗和を確率で加重和したもの

$$\sigma^2 = \mathrm{var}(Y) = E\left[(Y-\mu)^2\right] = \sum_{i=1}^{k} (y_i - \mu)^2 \Pr(Y = y_i)$$

- 標準偏差 standard deviation
  - 分散の平方根

期待値

- 期待値 expected value
  - 一般に, 確率を重みとする加重平均のことを期待値と呼ぶ
  - 平均:実現値の期待値
  - 分散:平均を引いたものの2乗の期待値
  - が, ふつうに「期待値」というときには平均をさす

共分散, 相関係数

- 共分散 covariance
  - 平均との差の積の期待値

$$\sigma_{XY} = \mathrm{cov}\left(X,Y\right) = E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]$$
$$= \sum_{i=1}^{k}\left(x_i-\mu_X\right)\left(y_i-\mu_Y\right)\Pr\left(X=x_i, Y=y_i\right)$$

  - 2つの変数が同じ方向に動く傾向があるとき正の値
  - 2つの変数が逆の方向に動く傾向があるとき負の値
- 相関係数 correlation coefficient
  - 共分散を標準偏差の積で割ったもの
  - 相関係数は-1より大きく, 1より小さい
  - 相関係数がゼロであるとき,「無相関」という
  - 2つの変数が独立であるとき, 無相関（逆は必ずしも成り立たない）

期待値の性質

- 期待値の線形性
  - 確率変数X, Y, 定数a, bに対して以下が成り立つ
  - $\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y]$
- 分散の性質
  - $\mathrm{var}(aX + bY) = a^2\mathrm{var}(X) + b^2\mathrm{var}(Y) + 2\,ab\,\mathrm{cov}(X, Y)$