応用システム工学
第五回　回帰分析

平成23年05月13日

チェビシェフの不等式

- 集合　$I = \left\{ x : \left| x - E[x] \right| \geq k\sqrt{V[x]} \right\}$　ただしk>0
- 確率変数Xがxとなる確率P(X=x)の確率頻度関数f(x)　$P\left[\left| X - E[X] \right| \geq k\sqrt{V[X]} \right] = \sum_{x \in I} f(x)$
  – 確率Pで分散を表す

$$V[X] \geq \sum_{x \in I} (x - E[X])^2 f(x)$$

$$\geq k^2 V[X] P\left[\left| X - E[X] \right| \geq k\sqrt{V[X]} \right]$$

$$\frac{1}{k^2} \geq P\left[\left| X - E[X] \right| \geq k\sqrt{V[X]} \right]$$

- 確率変数と期待値の差が分散の根より大きくなる確率は小さい($1/k^2$)

確率収束

- 確率変数として標本平均を考える
  - 期待値と分散 $E\left[\overline{X}_n\right]=\mu$ $V\left[\overline{X}_n\right]=\frac{1}{n}\sigma^2$
  - チェビシェフの不等式

$$\frac{1}{k^2} \geq P\left[\left|\overline{X}_n - E\left[\overline{X}_n\right]\right| \geq k\sqrt{V\left[\overline{X}_n\right]}\right] = P\left[\left|\overline{X}_n - \mu\right| \geq k\sqrt{\frac{\sigma^2}{n}}\right]$$

$\varepsilon = k\sqrt{\frac{\sigma^2}{n}}$ とすると $\frac{1}{k^2} = \frac{\sigma^2}{n\varepsilon^2} = P\left[\left|\overline{X}_n - \mu\right| \geq k\sqrt{\frac{\sigma^2}{n}}\right]$

$$\lim_{n\to\infty} P\left[\left|\overline{X}_n - \mu\right| \geq k\sqrt{\frac{\sigma^2}{n}}\right] = \lim_{n\to\infty}\frac{\sigma^2}{n\varepsilon^2} = 0$$

標本サイズnを大きくすると，標本平均と母平均に差がある確率が0となる

不変分散

- 一致推定量→標本サイズが十分大きいとき，標本平均が母平均に確率収束する性質をもつ推定量→標本平均は母平均に確率収束
- 不変推定量→期待値が母数に等しい推定量
  - 標本平均 $\overline{X}$ は不偏推定量
    - 母平均 μ に対して $E\left[\overline{X}\right]=\mu$

標本平均は
母平均と等しくなる

  - 標本の分散は不偏推定量となるか？
    - 標本の分散

$$S^2 = \frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \cdots + \left(X_n - \overline{X}\right)^2}{n}$$

不変分散

- 標本の分散

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[\left(X_i - \mu\right) - \left(\overline{X} - \mu\right)\right]^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \overline{Y}\right]^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - 2\frac{\overline{Y}}{n}\sum_{i=1}^{n}Y_i + \frac{\overline{Y}^2}{n}\sum_{i=1}^{n}1$$
$$= \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - 2\overline{Y}\overline{Y} + \frac{\overline{Y}^2 n}{n} = \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \overline{Y}^2$$

変数変換

$$Y_i = X_i - \mu$$
$$\overline{Y} = \overline{X} - \mu$$

不変分散

- 標本の分散の期待値

$$E\left[S^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \overline{Y}^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}Y_i^2\right] - E\left[\overline{Y}^2\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}E\left[Y_i^2\right] - E\left[\overline{Y}^2\right] = \frac{1}{n}\sum_{i=1}^{n}E\left[\left(X_i - \mu\right)^2\right] - E\left[\left(\overline{X} - \mu\right)^2\right]$$

– 標本サイズが大きいときの標本と母平均の差の二乗平均

$$E\left[\left(X_i - \mu\right)^2\right] \to \sigma^2$$

– 標本平均の分散

$$E\left[\left(\overline{X} - \mu\right)^2\right] = \frac{\sigma^2}{n}$$

不変分散

- 標本分散の期待値

$$E\left[S^2\right]=\frac{1}{n}\sum_{i=1}^{n}\sigma^2-\frac{\sigma^2}{n}$$

$$=\frac{1}{n}n\sigma^2-\frac{\sigma^2}{n}=\frac{n-1}{n}\sigma^2$$

  – 標本分散 S^2 の期待値$E\left[S^2\right]$は母分散 σ^2とは等しくない
  →標本分散S^2は母分散の不偏推定量ではない
  $\frac{n}{n-1}S^2$ が母分散の不偏推定量に相当

- 不偏分散（自由度n-1）

$$s^2=\frac{\left(X_1-\overline{X}\right)^2+\left(X_2-\overline{X}\right)^2+\cdots+\left(X_n-\overline{X}\right)^2}{n-1}$$

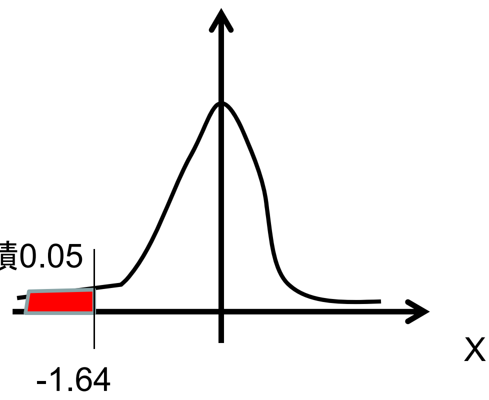
検定

- 仮説検定 → 統計的推測により母集団に対して判断する
- 帰無仮説 → 推論
  – 例 コイン投げの表が出る確率は50%
  – 帰無仮説を棄却する → 表が出る確率が50%というのは間違いである
- 対立仮説 → 帰無仮説が棄却された結果得られる結論
  – 表が出る確率は50%より小さい(大きい)

検定

- 有意水準　→　帰無仮説が正しい時に，帰無仮説を棄却する確率
  - 第一種の誤り　→　帰無仮説が正しい時に，帰無仮説を棄却する誤り
  - 第二種の誤り　→　対立仮説が正しい時に，帰無仮説を採用する誤り
    - 第一種の誤りの低減と，第二種の誤りの低減にはトレードオフの関係がある

検定



- 確率変数X
  - 標準正規分布N(0,1)
    - 期待値0，分散1
  - 発生する確率
    (有意水準)5%以下
    - X≦-1.64
    - Xが-1.64より小さければ，
      有意水準5%で帰無仮説が棄却される
  - 検出力　→　対立仮説が正しく，帰無仮説を棄却する確率(標本サイズが大きくなると検出力は大きくなる)

区間推定

- 正規分布する母集団の母平均を，標本平均を用いて推測
  - 区間推定　母平均がある区間に入っている確率を求める
    - 信頼係数　W→W x100%信頼区間
  - 母集団　平均μ，分散σ^2の正規分布N(μ, σ^2)
    - 確率密度関数 $f_x\left(x;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$
    - 標本(n個)も正規分布する $x_1, x_2, \cdots, x_n$
      - 標本平均も正規分布 $\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}$

正規分布の再生性
確率変数X,Yが独立で，各々正規分布となるときX+Yも正規分布となる

確率分布

- 確率変数Xの確率密度関数 $f(x)$
  - Xがaからbの値をとる確率

$$P[a \le x \le b]=\int_a^b f(x)dx$$

- 累積分布関数→確率変数Xがx以下の値である確率

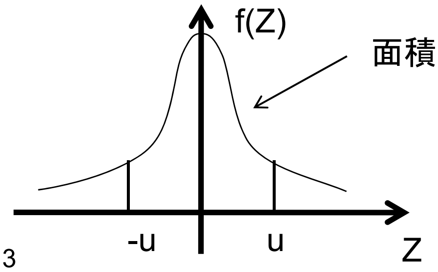
$$F[X]=P[X \le x]=\int_{-\infty}^x f(t)dt$$

区間推定

- 標本平均を正規分布化して考える
  X→Z $Z=\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$
  – 標本平均の分散 $\frac{\sigma^2}{n}$

$$P(-u \le Z \le u)=W$$

  – -u≦Z ≦uとなる確率W
    • 標本平均で表す



$$P\left(-u \le \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \le u\right)=W$$

区間推定

$$P\left(-u \le \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \le u\right)=W$$

- 信頼区間Wの上限と下限

$$P\left(-\overline{X}-u\sqrt{\frac{\sigma^2}{n}} \le -\mu \le -\overline{X}+u\sqrt{\frac{\sigma^2}{n}}\right)=W$$

$$P\left(\overline{X}-u\sqrt{\frac{\sigma^2}{n}} \le \mu \le \overline{X}+u\sqrt{\frac{\sigma^2}{n}}\right)=W$$

- 100W%信頼区間 $\left[\overline{X}-u\sqrt{\frac{\sigma^2}{n}}, \overline{X}+u\sqrt{\frac{\sigma^2}{n}}\right]$

信頼区間

- 復元抽出
  - 標本を取り出した後，標本を母集団に戻した後，改めて標本を抽出
    - 標本の値のデータが母集団で占める割合は変化しない
  - 母集団の大きさが，信頼区間の幅に影響しない
    - 標本サイズが信頼区間の幅に影響する
- 非復元抽出
  - 標本を取り出した後，標本を母集団に戻さない
    - 標本の値のデータが母集団で占める割合は変化する

X^2分布

- 母平均の区間推定 → 母分散σ^2を用いて，標本平均の分散σ^2/nを表す
  - 母分散の点推定 → 不変分散で表す
- 各々標準正規分布N(0,1)の独立な確率変数$Z_1, Z_2, \cdots, Z_n$ に対する次式で定義される確率変数X^2

$$\chi^2 = Z_1{}^2 + Z_2{}^2 + \cdots + Z_n{}^2 = \sum_{i=1}^{n} Z_i{}^2$$

の確率分布→自由度nのX^2分布 $\chi^2(n)$

X^2分布

- 母分散σ^2を不偏分散s^2とX^2分布で表す
  - 母集団 → 正規分布N(μ, σ^2)
    - 標本サイズn個 $Z_1, Z_2, \cdots, Z_n$
      - 標本平均 $\overline{Z}$，不偏分散 s^2

$$s^2 = \frac{1}{n-1}\left\{\left(Z_1 - \overline{Z}\right)^2 + \left(Z_2 - \overline{Z}\right)^2 + \cdots + \left(Z_n - \overline{Z}\right)^2\right\}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(Z_i - \mu\right) - \left(\overline{Z} - \mu\right)\right]^2$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(Z_i - \mu\right)^2 - 2\left(\overline{Z} - \mu\right)\sum_{i=1}^{n}\left(Z_i - \mu\right) + \left(\overline{Z} - \mu\right)^2\sum_{i=1}^{n}1\right\}$$

X^2分布

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(Z_i - \mu\right)^2 - 2\left(\overline{Z} - \mu\right)n\left(\overline{Z} - \mu\right) + \left(\overline{Z} - \mu\right)^2 n\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(Z_i - \mu\right)^2 - n\left(\overline{Z} - \mu\right)^2\right\}$$

$$(n-1)s^2 = \sum_{i=1}^{n}\left(Z_i - \mu\right)^2 - n\left(\overline{Z} - \mu\right)^2$$

$$\frac{(n-1)s^2}{\sigma^2} + \left(\frac{\overline{Z} - \mu}{\sqrt{\sigma^2/n}}\right)^2 = \sum_{i=1}^{n}\left(\frac{Z_i - \mu}{\sigma}\right)^2$$

X^2分布

$$\frac{(n-1)s^2}{\sigma^2}+\left(\frac{\bar{Z}-\mu}{\sqrt{\sigma^2/n}}\right)^2=\sum_{i=1}^{n}\left(\frac{Z_i-\mu}{\sigma}\right)^2$$

$$\sum_{i=1}^{n}\left(\frac{Z_i-\mu}{\sigma}\right)^2$$ ➡ 自由度nのx²分布

$$\left(\frac{\bar{Z}-\mu}{\sqrt{\sigma^2/n}}\right)^2$$ ➡ $\bar{Z}$ の確率分布は $N\left(\mu,\frac{\sigma^2}{n}\right)$ より，自由度1のx²分布

$$\frac{(n-1)s^2}{\sigma^2}$$ ➡ 自由度n-1のx²分布となる