統計学II

早稲田大学政治経済学術院
西郷 浩

本日の目標

- 区間推定
  - 信頼係数(信頼水準)
- 母平均の区間推定
- 母比率(成功確率)の区間推定

利用データについて

- JGSS2010年
  - 抽出実験に用いたデータは、東京大学社会科学研究所付属社会調査・データアーカイブ研究センターSSJデータアーカイブのリモート集計システムを利用し、同データアーカイブが所蔵する[JGSS2010]の個票をデータを集計したものである。

区間推定

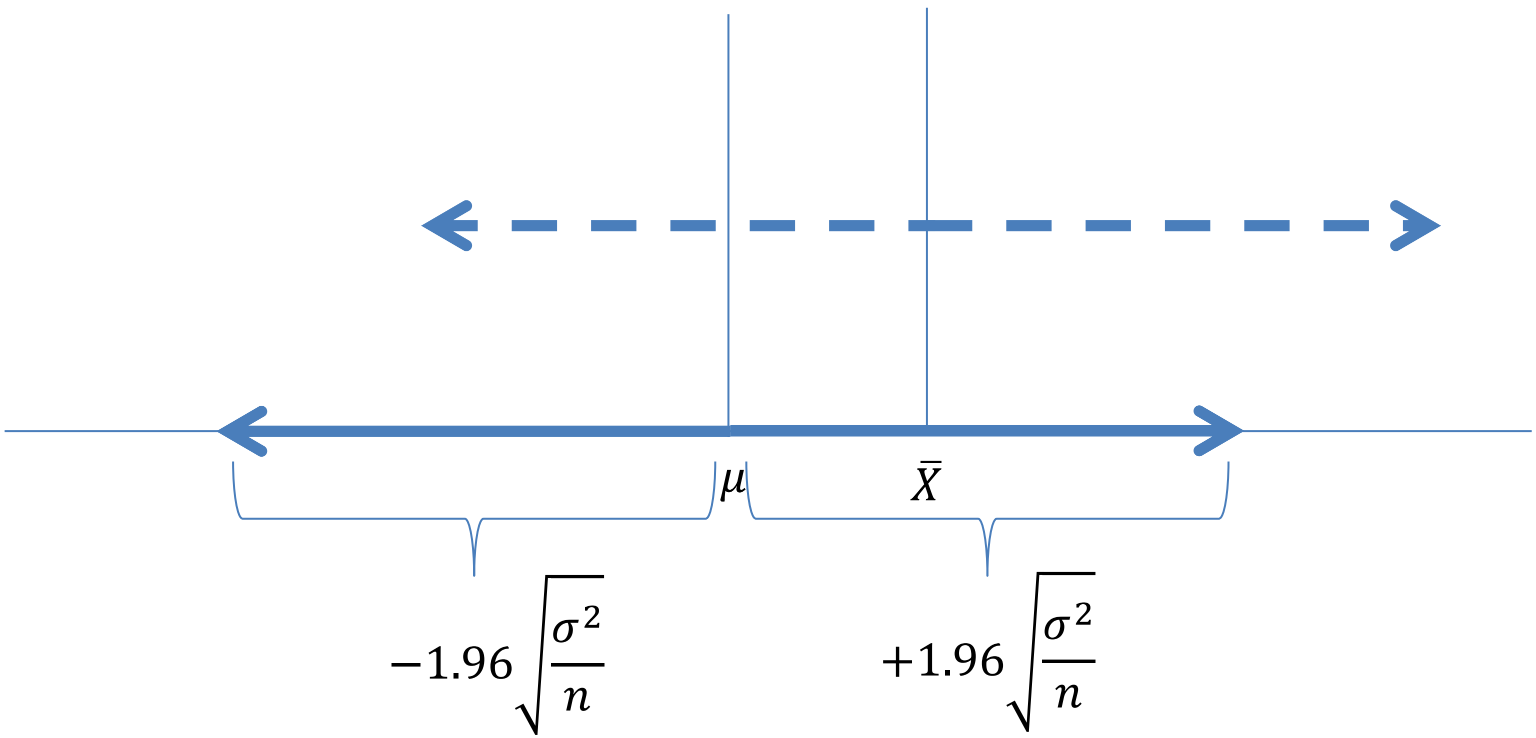
- 区間推定
  - 所与の水準以上の確率で母数を含む区間を構成する。
  - 例:
    - 母平均の区間推定
      - $P(L \leq \mu \leq U) \geq 0.95$
        - » L: 下限(標本から求める確率変数)
        - » U: 上限(標本から求める確率変数)
      - このような条件を満たす区間 $[L, U]$ を信頼係数0.95の信頼区間(区間推定値)とよぶ。
- 目標
  - 信頼係数を高く保ちつつ、幅の狭い区間を構成する。

母平均の区間推定(1)

- 信頼区間の構成
  - 標本：$X_i \sim_{iid} (\mu, \sigma^2)$
  - 中心極限定理：$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \to_d \ N(0,1)$
  - 中心極限定理から導かれる性質
    - $P\left(-1.96 \leq \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \leq 1.96\right) \approx 0.95$
  - カッコ内を同値（内容が同等）の式に変形する。
    - $P\left(\mu - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \bar{X} \leq \mu + 1.96\sqrt{\frac{\sigma^2}{n}}\right) \approx 0.95$
    - $P\left(\bar{X} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + 1.96\sqrt{\frac{\sigma^2}{n}}\right) \approx 0.95$
      - 信頼区間の構成まであともう一歩。

母平均の区間推定(2)

図1：$\mu \pm 1.96\sqrt{\sigma^2/n}$ と $\bar{X} \pm 1.96\sqrt{\sigma^2/n}$ との関係



母平均の区間推定(3)

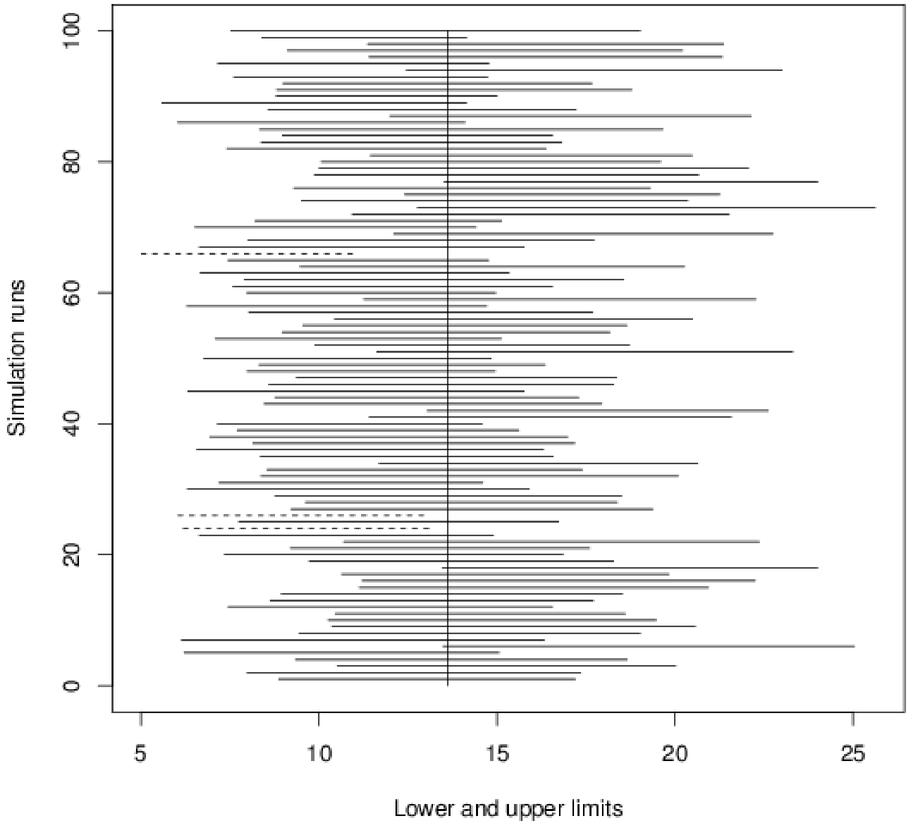
- 未知母数 σ^2 の処理
  - 標本推定量（標本不偏分散）で置き換える。
    - $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
      - 不偏性 $E(S^2) = \sigma^2$ をもつ。
      - $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ とすると、$E(\tilde{\sigma}^2) = \frac{n-1}{n}\sigma^2 < \sigma^2$ となり、平均的に過小推定。
  - 標本の大きさ n が大きければ、母分散 σ^2 を標本不偏分散で置き換えても、捕捉確率がそれほど変わらないことが知られている。
    - $P\left(\bar{X} - 1.96\sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{X} + 1.96\sqrt{\frac{S^2}{n}}\right) \approx 0.95$
- 信頼係数0.95の母平均 μ についての信頼区間（区間推定値）
  - $\left[\bar{X} - 1.96\sqrt{S^2/n}, \bar{X} + 1.96\sqrt{S^2/n}\right]$

母平均の区間推定(4)

- 例:
  - 就労年数データ(p. 128)$\mu = 13.6$
    - 大きさ$n = 30$ の標本を抽出
    - 4, 7, 7, 1, 10, 1, 25, 0, 6, 4, 2, 5, 15, 6, 13, 2, 38, 2, 12, 2, 20, 12, 35, 18, 6, 6, 9, 20, 5, 15.
    - 信頼係数0.95の信頼区間
      - $\bar{X}_{obs} \approx 10.3$, $S^2_{obs} \approx 92.6$
      - $\left[10.3 - 1.96\sqrt{92.6/30}, 10.3 + 1.96\sqrt{92.6/30}\right] = [6.8, 13.7]$
- 実験
  - 標本の大きさを n = 30 として標本抽出を100,000回繰り返した結果:
    - 経験的な捕捉率:0.93(信頼係数つまり名目捕捉率である0.95 に近い)
  - 最初の100回の結果(次のスライド)

母平均の区間推定(5)

図2：就労年数データによる区間推定値の実験結果の一部



成功確率（母比率）の区間推定(1)

- ベルヌーイ確率変数
  - $X_i \sim_{iid} Bernoulli(p)\ (i = 1, 2, \ldots, n)$
    - $X_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$
  - $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{p}$
  - $E(\hat{p}) = E(X_i) = p$
  - $V(\hat{p}) = \frac{V(X_i)}{n} = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$
  - σ^2の不偏推定量：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$
- 信頼係数0.95の成功確率 p についての信頼区間
  - $\left[\bar{X} - 1.96\sqrt{S^2/n}, \bar{X} + 1.96\sqrt{S^2/n}\right] =$
    $\left[\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/(n-1)}, \hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/(n-1)}\right]$

成功確率（母比率）の区間推定(2)

- 例：
  - JGSS2010における「現在の仕事の満足度」データ $p = 0.664$
    - 50人無作為抽出したうち、35人が「満足」または「どちらかといえば満足と回答した。
    - 信頼係数0.95の信頼区間
      - $\hat{p}_{obs} = \frac{35}{50} = 0.7, S_{obs}^2 = \frac{50 \times 0.7(1-0.7)}{50-1} \approx 0.214$
      - $\left[0.7 - 1.96 \times \sqrt{\frac{0.7(1-0.7)}{50-1}}, 0.7 + 1.96 \times \sqrt{\frac{0.7(1-0.7)}{50-1}}\right] = [0.57, 0.83]$
- 実験
  - 標本の大きさを n = 50 として標本抽出を100,000回繰り返した結果：
    - 経験的な捕捉率：0.92（信頼係数つまり名目捕捉率である0.95 に近い）

成功確率（母比率）の区間推定(3)

- 推定精度に応じた標本の大きさの決定
  - 「満足」または「どちらかといえば満足」と回答する人の母集団比率を、±0.02（±2%）の精度で推定したい。
  - 中心極限定理により、近似的に以下の性質が成り立つ。
    - $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$
  - したがって以下の式が成り立つ。
    - $P\left(-1.96\sqrt{p(1-p)/n} \leq \hat{p}-p \leq 1.96\sqrt{p(1-p)/n}\right) \approx 0.95$
  - 推定精度±0.02を達成するには以下の条件が満たされればよい。
    - $1.96\sqrt{p(1-p)/n} \leq 0.02$
  - 2つの場合
    - 母比率 p について見当がつく：上の不等式を n について解く。
    - 母比率 p について見当がつかない：最悪の場合として $p=1/2$ を想定し、上の不等式を n について解く。
      - $1.96 \approx 2$ とみなせば、$1/\sqrt{n} \leq 0.02 \Leftrightarrow n \geq 1/0.02^2 = 2500$

正規母集団からの標本にもとづく区間推定 (1)

- 標本の大きさ n が小さい。
  - 正規分布にもとづく区間推定の精度が悪化する。
    - 中心極限定理に依拠する正規分布による近似の精度が低い。
    - 母分散 σ^2 の推定量 S^2 の推定精度が低い。
  - しかし、「母集団が正規分布にしたがう」ときには、簡単な解決策がある。
- 正規母集団からの標本にもとづく区間推定
  - 標本：$X_i \sim_{iid} N(\mu, \sigma^2)$ $(i = 1, 2, \dots, n)$
  - このとき、以下の性質が成り立つ。
    - $\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t(n-1)$　参考：$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$
  - これを利用すれば、以下の性質が導かれる。
    - $P\left\{-t_{0.025}(n-1) \le \frac{\bar{X}-\mu}{\sqrt{S^2/n}} \le t_{0.025}(n-1)\right\} = 0.95$
      - ただし、$t_{0.025}(n-1)$ は自由度 n-1 のt分布の上側0.025（2.5%）点

正規母集団からの標本にもとづく区間推定 (2)

- 信頼係数0.95の母平均 μ の信頼区間（正規母集団の場合）
  - $\left[\bar{X} - t_{0.025}(n-1)\sqrt{S^2/n}, \bar{X} + t_{0.025}(n-1)\sqrt{S^2/n}\right]$

- 例:
  - NB10（名目的な重量が10gの重り）の測定結果
    - 9.0999591, 9.999600, 9.999594, 9.999601, 9.999598
  - 正規母集団を仮定する。
    - 測定誤差であることと、過去の経験とから妥当な仮定とみなせる。
  - 信頼係数0.95の信頼区間
    - $\bar{X}_{obs} = 9.999597, S_{obs}^2 = 1.77 \times 10^{-11}, t_{0.025}(5-1) = 2.776$
    - $\left[9.999579 - 2.776\sqrt{\frac{1.77\times10^{-11}}{5}}, 9.999579 + 2.776\sqrt{\frac{1.77\times10^{-11}}{5}}\right] = [9.99592, 9.99602]$

正規母集団からの標本にもとづく区間推定(3)

表1：t分布表の見方

ν		.025 (.050)	
…		…	
4	…	2.776	…
…		…	

- t分布表の利用方法
  - 教科書付表B（p. 225）
    - 表側（表の左側）：自由度 ν
    - 表頭（表の上側）：右側確率
      - カッコ内はその倍＝両側確率 を示す。
        - » 信頼係数 0.95 の信頼区間を構成するときには、$\frac{1-0.95}{2}=0.025$ （カッコ内は0.05）に対応する部分を見る。
    - 例：信頼係数0.95 で自由度 $5-1=4$ のとき。
      - 右上の表1参照