母比率の区間推定

樋口さぶろお https://hig3.net

龍谷大学 先端理工学部 数理・情報科学課程

確率統計 I L12(2023-06-26 Mon)

最終更新: Time-stamp: ”2024-06-28 Fri 17:08 JST hig”

今日の目標

- 母比率を区間推定できる 岩薩林 確率・統計 §7.3
- カイ二乗分布を説明できる 岩薩林 確率・統計 p.123

L10-Q3

Quiz 解答: ベルヌーイ分布の独立同分布の和と中心極限定理

100 回中表の出る回数は，$Y = X_1 + \cdots + X_{100}$, $X_i \sim \mathrm{B}(1, \frac{4}{5})$, 独立同分布, $\mathrm{E}[X_i] = \frac{4}{5}, \mathrm{V}[X_i] = \frac{4}{5}\frac{1}{5}$. よって, $\mathrm{E}[Y] = 80$, $\mathrm{V}[Y] = 4^2$ である (これは, $Y \sim \mathrm{B}(100, \frac{4}{5})$ からも求められる)

$n = 100$ が大きいと考えると, 中心極限定理より, Y は近似的に正規分布 $\mathrm{N}(80, 4^2)$ にしたがう.

標準化された $Z = \frac{Y-80}{4}$ は近似的に標準正規分布 $\mathrm{N}(0, 1^2)$ にしたがう.

よって, 求める確率は,

$P(73 < X \le 79) = P(-\frac{7}{4} < Z \le -\frac{1}{7}) = \Phi(-\frac{1}{4}) - \Phi(\frac{7}{4}) =$
$0.4599 - 0.0987 = 0.3612.$

L11-Q1

L11-Q2

Quiz 解答: 母平均値, 母分散, 母比率の点推定

在庫のフライドチキンの重さを X とすると,

1. 標本平均値 (の実現値) は $\overline{X} = \frac{1}{6}(117 + \cdots + 112) = 111\mathrm{g}$ なので, 母平均値は $111\mathrm{g}$ と推定できる.
2. 標本期待値 (の実現値) は $\overline{X^2} = \frac{1}{6}(117^2 \cdots + 112^2) = 37078/3\mathrm{g}^2$ なので, 母期待値は $37078/3\mathrm{g}^2$ と推定できる.
3. 不偏標本分散 (の実現値) は,
$S^2 = \frac{1}{6-1}[(117-111)^2 + \cdots + (112-111)^2] = 46\mathrm{g}^2$ なので, 母分散は $46\mathrm{g}^2$ と推定できる.
4. 標本比率 (の実現値) は, $p = \frac{1}{6}[1+0+0+1+0+1] = 0.5$ なので, 母比率は 0.5 と推定できる.

ここまで来たよ

点推定 対 区間推定

点推定 岩薩林 確率・統計 §6.1

真の母平均値はわからないが, 標本平均値を使って,

「母平均値を A 円と推定する」

それどのくらい正確なの? 正確さは実は 母分散や標本サイズによる

区間推定 岩薩林 確率・統計 §6.1

「母平均値が, B 円以上 C 円以下である '確率' は $1-\alpha=0.95$」

推定の精度・正確さまで表現

ここで '確率' というのは不誠実. 正しい言葉遣いは, 信頼係数=信頼度で

「母平均値の信頼係数 $1-\alpha=0.95$ の信頼区間は B 円以上 C 円以下」

動く (確率変数である) のは母平均値 μ でなく, B,C のほう.

標本平均値のしたがう分布 (正規母集団, 母分散既知)

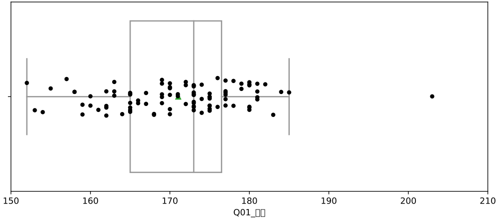
岩薩林 確率・統計 p.144

$\mathrm{N}(\mu, \sigma^2)$ にしたがう母集団 (正規母集団) の, サイズ n の標本を何回も取り出して, 毎回, 標本平均値 $\overline{X}$ を計算する.
その意味で, 標本平均値は確率変数
先週チーム課題で計算したのは, 1 標本で得られた標本平均値 (1 試行で得られた確率変数の値).
実は, $U_n = \overline{X} \sim \mathrm{N}(\mu, \sigma^2/n)$. $Z = \frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} \sim \mathrm{N}(0, 1^2)$.

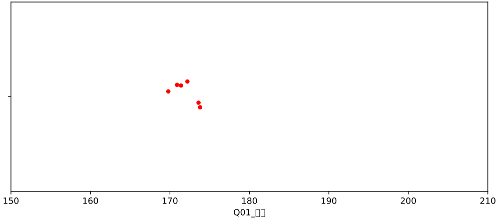
正規分布の再生性 確率統計 II(2023)L?? 岩薩林 確率・統計 p.99 から. 使わなくても, $n \to +\infty$ で正しいことは中心極限定理からわかる. 正規母集団でないときも, 標本サイズ n が大きい (30 くらい) なら, 近似的に成立することが多い.

身長の標本平均値の分布と母平均値

母集団



サイズ 10 の標本 6 個の標本平均値

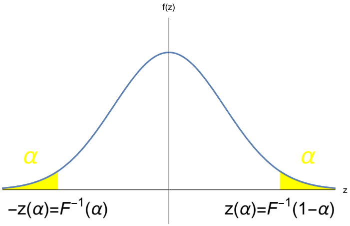


L12-Q1

Quiz(正規分布の上側確率)

$Z \sim \mathrm{N}(0, 1^2)$ とする. Z の累積分布関数を $\Phi(z)$ とする.

1. $P(u < Z)$ を u と Φ で表そう.
2. $P(u < Z) = \alpha$ となる u を α と Φ で表そう. この u のことをよく $z(\alpha)$ と書く.
3. Z の確率密度関数が偶関数であることから, $P(u < Z) = P(Z < -u)$ であることを説明しよう.
4. $P(Z < -u) + P(u < Z) = \alpha$ となる u を α と $z()$ で表そう.
5. $\alpha = 0.05$ のとき, 上のような u を求めよう.



定義 (標準正規分布の $z(\alpha)$)

$Z \sim \mathrm{N}(0, 1^2)$ のとき，上側確率 $P(Z \geq z(\alpha)) = \alpha$ となる境い目を $z(\alpha)$ と定める． 岩薩林 確率・統計 付表 1 下

$z(\alpha) = F_Z^{-1}(1-\alpha)$．偶関数だから $z(1-\frac{\alpha}{2}) = -z(\frac{\alpha}{2})$．

標本平均値 $\overline{X}$ が母平均値 μ から大きく外れない確率は大きい (ここでは $1-\alpha = 1-0.05$) という式を書くと…

$$P\left(F_{\mathrm{Z}}^{-1}(\tfrac{\alpha}{2}) < \frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} < F_{\mathrm{Z}}^{-1}(1-\tfrac{\alpha}{2})\right) = 1-\alpha.$$

$$P\left(\mu + F_{\mathrm{Z}}^{-1}(\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n} < \overline{X} < \mu + F_{\mathrm{Z}}^{-1}(1-\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n}\right) = 1-\alpha.$$

μ について不等式を解くと,

$$P\left(\overline{X} - F_{\mathrm{Z}}^{-1}(1-\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n} < \mu < \overline{X} - F_{\mathrm{Z}}^{-1}(\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n}\right) =1-\alpha.$$

$$P\left(\overline{X} - z(\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n} < \mu < \overline{X} - (-z(\tfrac{\alpha}{2})) \times \sqrt{\sigma^2/n}\right) =1-\alpha.$$

業界の習慣で, しばしば $\alpha = 0.05, 0.01$

$$\alpha = 0.05 \rightsquigarrow P\left(\overline{X} - 1.96 \times \sqrt{\sigma^2/n} < \mu < \overline{X} + 1.96 \times \sqrt{\sigma^2/n}\right) =1-0.05.$$

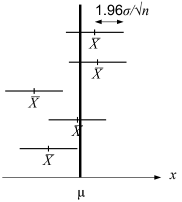
命題 (母平均値 (正規母集団, 母分散既知) の信頼区間 岩薩林 確率・統計 定理 6.1(p.18))

$\mathrm{N}(\mu, \sigma^2)$ にしたがう母集団の, σ^2 がわかっているとき, サイズ n の標本から区間推定すると, 母平均値 μ の 信頼係数 $1-\alpha$ の信頼区間 ($(1-\alpha)$ 信頼区間) は, $\overline{X}$ を標本平均値として,

$$\overline{X} - z(\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n} < \mu < \overline{X} + z(\tfrac{\alpha}{2}) \times \sqrt{\sigma^2/n}.$$

何回も標本抽出して何個も信頼区間を求めたとき, 信頼区間が μ を含む確率は, 信頼係数 $1-\alpha$. 推定が外れる確率 α.

切りがいい α の $z(\alpha)$ は岩薩林 確率・統計 付表 1 下 (p.227)
$z(\frac{0.05}{2}) = 1.96, z(\frac{0.01}{2}) = 2.58.$



ゆるゆるな高校 数学 B では, $z(\frac{0.05}{2}) = 1.96$ の場合のみ.
$a < \mu < b$ でなく, 閉区間の記号 $[a, b]$ で.
真の母分散 σ^2 の代わりに, (不偏 $\frac{1}{n-1}$ でない)
標本分散 $S^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$ を使っていい.

L12-Q2

Quiz(母平均値の区間推定 (母分散既知))

確率変数 X は, 母分散が $\mathrm{V}[X]=\sigma^2=3^2$, 母平均値 $\mu=\mathrm{E}[X]$ が不明な正規分布 $\mathrm{N}(\mu,3^2)$ にしたがう. X のサイズ 4 の標本を抽出したところ次のようになった (本当は整数値になる確率は 0 だけど, 計算を楽にするためのファンタジー).
$51, 52, 47, 50.$

1. 母平均値 μ を, 信頼係数 $1-\alpha=0.95$ で区間推定しよう.
2. 母平均値 μ を, 信頼係数 $1-\alpha=0.99$ で区間推定しよう.

岩薩林 確率・統計 例題 6.4(p. 145) | 岩薩林 確率・統計 問題 3(p.146) | 岩薩林 確率・統計 第 6 章練習問題 1

推定が正確であるとは 信頼区間が 自分の言葉で であること.

Quiz(区間推定の性質)

標本からの母平均値の区間推定について, 正しいのはどれ?

1. 母分散が大きいほど, 信頼区間は大きくなる
2. 標本サイズが大きいほど, 信頼区間は大きくなる
3. 母平均値が大きいほど, 信頼区間は小さくなる
4. 信頼係数が大きいほど, 信頼区間は小さくなる

標本平均値が大きい $\Rightarrow$ 信頼区間は 平行移動する

母分散が大きい $\Rightarrow$ 信頼区間は 大きい

標本サイズ n が小さい 信頼区間は 小さい

ここまで来たよ

母比率の信頼区間

母比率は母平均値の一種なので，さっきの区間推定の式で，$\sigma^2 = p(1-p)$ とおく．
信頼係数 $1-\alpha$．

$$P\left(p - z(\tfrac{\alpha}{2}) \times \sqrt{\tfrac{p(1-p)}{n}} < \hat{p} < p + z(\tfrac{\alpha}{2}) \times \sqrt{\tfrac{p(1-p)}{n}}\right) = 1-\alpha.$$

平方根の中の $\sigma^2 = p(1-p)$ は $\hat{p}(1-\hat{p})$ と近似して p について解くと次を得る．

母比率の信頼区間 (母分散未知) 岩薩林 確率・統計 §7.3

X のサイズ n の標本で，標本比率 $\hat{p} = k/n$ のとき，母比率の 信頼係数 $1-\alpha$ の信頼区間 ($(1-\alpha)$ 信頼区間) は，

$$\hat{p} - z(\tfrac{\alpha}{2}) \times \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z(\tfrac{\alpha}{2}) \times \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}}.$$

$z(\frac{0.05}{2}) = 1.96, z(\frac{0.01}{2}) = 2.58.$

L12-Q3

Quiz(母比率の区間推定)

選挙で出口調査をしたところ, 50 人中 35 人が A 候補に投票したと答えた. 母集団を投票した人 2500 人とする. そのうち A 候補に投票した人の母比率 (得票率) を考える.

1. A 候補の得票率を, (点) 推定しよう
2. A 候補の得票率を, 信頼係数 $1-\alpha=0.95$ で区間推定しよう.
3. A 候補の得票率を, 信頼係数 $1-\alpha=0.99$ で区間推定しよう.

岩薩林 確率・統計 例題 7.6(p.170) | 岩薩林 確率・統計 問題 7(p.171) | 岩薩林 確率・統計 第 7 章練習問題 2(2)

注: 下限, 上限が 0,1 を越えるときは, 0,1 に直してしまっていい.

ここまで来たよ

11 母集団と標本・母平均値/母分散/母比率の点推定

12 母比率の区間推定
- 母平均値の区間推定 (正規母集団, 母分散既知)
- 母比率の区間推定
- 不偏標本分散のちらばりとカイ二乗分布
- 母分散の区間推定

ちらばり (不偏標本分散) のちらばりを考えたい

標本データのちらばりって? $\sqrt{母分散} \overset{点推定}{\longleftarrow} \sqrt{不偏標本分散}$

- 母分散の点推定の精度って?

	の点推定	の区間推定
母平均値 μ	標本平均値 $\overline{X} = \frac{1}{n}[X_1 + \cdots]$	$\overline{X} - \boxed{正}\sqrt{\quad} < \mu < \overline{X} - \boxed{負}\sqrt{\quad}$
母分散 σ^2	不偏標本分散 $S^2 = \frac{1}{n-1}[(X_1 - \overline{X})^2 + \cdots]$	$S^2 \times \boxed{小} < \sigma^2 < S^2 \times \boxed{大}$

母集団が正規分布にしたがうとき

- 標本平均値の分布 (正規分布) をうまく平行移動, 拡大縮小すると標準正規分布 $\mathrm{N}(0, 1^2)$
- 不偏標本分散の分布をうまく拡大縮小するとカイ二乗分布 χ_k^2

確率統計 II(2023)L??

カイ二乗分布 岩薩林 確率・統計 p.123

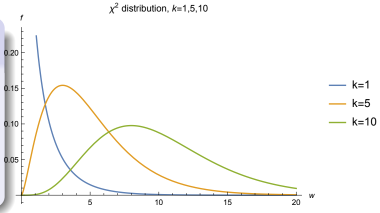
カイ二乗分布

$Z_1, \ldots, Z_k \sim \mathrm{N}(0, 1^2)$, iid のとき, 確率変数 $W = Z_1^2 + \cdots + Z_k^2$ は, 自由度 k のカイ二乗分布 χ_k^2 にしたがう.

言語	小	大	読み
英語	x	X	エクス
ギリシャ語	χ	X	カイ

χ_k^2 の確率密度関数

$$f_k(y) = \begin{cases} C_k \times y^{\frac{k}{2}-1} \mathrm{e}^{-\frac{1}{2}y} & (y \geq 0) \\ 0 & (\text{他}) \end{cases}$$



$W_k \sim \chi_k^2$ に対して,

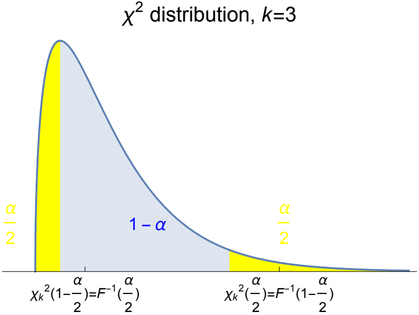
$\mathrm{E}[W_k] = \mathrm{E}[Z_1^2 + \cdots + Z_k^2] = k, \mathrm{V}[W_k] = 2k, \mathrm{E}[(W_k)^\ell] =$ 簡単じゃない.

カイ二乗分布の確率密度関数と累積分布関数

```
rvx2=stats.chi2(df=\text{自由度}k)
```

上側確率 $\alpha = P(W > \chi_k^2(\alpha))$ となる境い目 $\chi_k^2(\alpha) = F^{-1}(1-\alpha)$.

岩薩林 確率・統計 付表 3　付表 1 とフォーマットが違う



$\chi_k^2(\alpha)$ の定義

岩薩林 確率・統計 例題 5.7

$\alpha = P(W > \chi_k^2(\alpha))$.

$\chi_k^2(\alpha) = F^{-1}(1-\alpha)$.

L12-Q4

Quiz(カイ二乗分布の確率と $\chi^2_k(\alpha)$)

標準正規分布にしたがう確率変数 $Z \sim \mathrm{N}(0, 1^2)$ と，自由度 $k = 1$ のカイ二乗分布にしたがう確率変数 $W \sim \chi^2_1$ を考える．各分布の累積分布関数の逆関数 Φ^{-1}, F^{-1} を使って答え，さらに Python, 数表を使って数値にしよう．

1. 確率 $P(Z > x_0) = 0.025$ となる $x_0 = z(0.025)$ を求めよう．
2. 確率 $P(Z > x_0) = 1 - 0.025$ となる $x_0 = z(1 - 0.025)$ を求めよう．
3. 確率 $P(W > w_0) = 0.05$ となる $w_0 = \chi^2_1(0.05)$ を求めよう．
4. 確率 $P(W > w_0) = 1 - 0.05$ となる $w_0 = \chi^2_1(1 - 0.05)$ を求めよう．

ここまで来たよ

11 母集団と標本・母平均値/母分散/母比率の点推定

12 母比率の区間推定
- 母平均値の区間推定 (正規母集団, 母分散既知)
- 母比率の区間推定
- 不偏標本分散のちらばりとカイ二乗分布
- 母分散の区間推定

不偏標本分散のしたがう分布

不偏標本分散のしたがう分布 岩薩林 確率・統計 定理 5.6

正規分布 $\mathrm{N}(\mu, \sigma^2)$ からサイズ n の標本 $X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, iid を取り出すとき, 不偏標本分散

$$S^2 = \frac{1}{n-1}[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]$$

から定めた

$$W = (n-1) \times \frac{S^2}{\sigma^2}$$

は, 自由度 $k = n - 1$ の カイ二乗分布 χ^2_{n-1} にしたがう.

比 $\frac{\text{不偏標本分散}}{\text{母分散}}$ は 1 に近いところに分布するが, 実は, 確率変数 $\frac{W}{n-1}$. ($W \sim \chi^2_{n-1}$)

証明じゃないけど説明

独立な $X_i \sim \mathrm{N}(\mu, \sigma^2)$ $(i = 1, \dots, n)$ に対して,

$$n \times \frac{1}{n}\left[\left(\frac{X_1 - \mu}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \mu}{\sigma}\right)^2\right]$$

は自由度 n のカイ二乗分布 χ^2_n にしたがう.
不偏標本分散 S^2 に対して,

$$W = (n-1) \times \frac{S^2}{\sigma^2} = (n-1) \times \frac{1}{n-1}\left[\left(\frac{X_1 - \overline{X}}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \overline{X}}{\sigma}\right)^2\right]$$

は自由度は $n-1$ のカイ二乗分布 χ^2_{n-1} にしたがう.
$-\mu$ でなく $-\overline{X}$ であるため自由度 $n-1$.