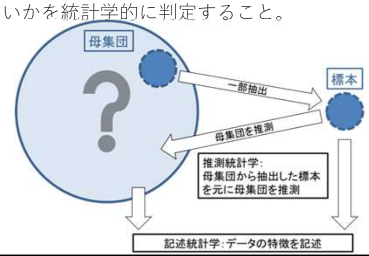
推定

佐賀大学　新井康平

1

推定：母集団を特徴づける母数（パラメーター：平均など）を統計学的に推測すること。
検定：母集団から抽出された標本の統計量に関する仮説が正しいかを統計学的に判定すること。



2

点・区間推定

- 日本人全員（母集団）からランダムに100人を抽出した標本を例に取ると、
- 推定：抽出された100人の身長から日本人全員の平均身長を推測すること。
  - 点推定：平均値などを1つの値で推定すること。
  - 区間推定：平均値などをある区間で推定すること。
- 検定：日本人の平均身長が165cmと言われているが、抽出された100人の平均は167cmだった。このとき、100人の平均値は妥当かどうかを判定する

3

点推定

- 10都道府県にある映画館の合計スクリーン数の平均は「93.8」でした。
- この「93.8」を47都道府県（＝母集団）の平均値（つまり母平均）と見なしてしまう、とするのが母平均の点推定

No.	都道府県	全スクリーン数
1	兵庫	126
2	大阪	224
3	奈良	34
4	岩手	25
5	千葉	199
6	茨城	89
7	福岡	178
8	山梨	14
9	滋賀	38
10	鳥取	11

4

標本分散

- 標本分散は一致推定量ではあるものの不偏推定量ではありません。つまり、nが十分に大きくない場合には標本分散の期待値は母分散に一致せず、母分散より小さくなります。
- 標本分散の期待値が母分散に一致するように標本分散の算出式にn/(n-1)をかけたものが不偏分散の算出式となります。したがって、不偏分散は一致性と不偏性をもつ推定量です

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad s^2 = \frac{n}{n-1}\times\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

5

無作為に抽出された10都道府県の合計スクリーン数のデータから不偏分散

- 不偏分散
- 不偏標準偏差

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{10-1}\{(126-93.8)^2 + (224-93.8)^2 + \cdots + (11-93.8)^2\} \\ &= 6757.3 \end{aligned}$$

6

標準誤差

$$SE = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}}{\sqrt{n}}$$

- 標本平均の標準偏差

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

- 標本平均の不偏標準偏差（＝標準誤差SE）

$$SE = \frac{s}{\sqrt{10}} = \frac{82.2}{\sqrt{10}} = 26.0$$

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \sqrt{\frac{1}{10-1}\times\{(126-93.8)^2 + (224-93.8)^2 + \cdots + (11-93.8)^2\}} \\ &= \sqrt{6757.3} \\ &= 82.2 \end{aligned}$$

7

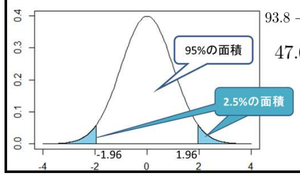
95%信頼区間←区間推定

- 「95%信頼区間」とは、母平均が95%の確率でその範囲にある

$$\overline{x} - 1.96\times\sqrt{\frac{\sigma^2}{n}} \le \mu \le \overline{x} + 1.96\times\sqrt{\frac{\sigma^2}{n}}$$

$$93.8 - 1.96\times\sqrt{\frac{5560}{10}} \le \mu \le 93.8 + 1.96\times\sqrt{\frac{5560}{10}}$$

$$47.6 \le \mu \le 140.0$$



8

区間推定

- 通常母平均の区間推定を行う場合にはt分布を用いた方法が使われます。
- 母平均の区間推定では「95%信頼区間（95%CI）」を求めることが多いですが、これは「母集団から標本を取ってきて、その標本平均から母平均の95%信頼区間を求める、という作業を100回やったときに、95回はその区間の中に母平均が含まれる」ということを意味します。言い換えると、「母集団から標本を取ってきて標本平均を求める、という作業を100回やったときに、95回はその標本平均を含んでいると考えられる区間のこと」

9

サンプル(標本)数

- サンプル数=100

$$93.8 - 1.96 \times \sqrt{\frac{5560}{100}} \le \mu \le 93.8 + 1.96 \times \sqrt{\frac{5560}{100}}$$

$$79.2 \le \mu \le 108.4$$

- サンプル数=1000

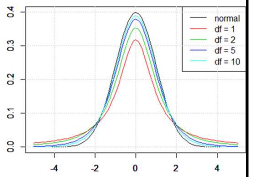
$$93.8 - 1.96 \times \sqrt{\frac{5560}{1000}} \le \mu \le 93.8 + 1.96 \times \sqrt{\frac{5560}{1000}}$$

$$89.2 \le \mu \le 98.4$$

10

母分散が分からない場合にはt分布を使って区間推定

- t分布は標準正規分布とよく似た形の分布で、パラメータである「自由度」によって分布の形が変わるという特徴を持っています。
- 自由度を変化させた時のt分布の形を見てみます。次のグラフは自由度（グラフ中ではdfで表示しています）が1, 2, 5, 10である場合のt分布（赤、緑、青、水色線）と標準正規分布（黒線：normal）を表したもの



11

t分布

- 平均μ、不偏分散s^2の正規分布に従う母集団から抽出したサンプルサイズnの標本を使って算出される次に示す統計量tの値は、自由度(n-1)のt分布に従います。

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

- 確率変数Xが自由度mのt分布に従っている時、Xの期待値と分散

$$E(X) = 0 \quad (m > 1)$$

$$V(X) = \frac{m}{m-2} \quad (m > 2)$$

12

不偏分散推定

- 指定した数値を正規母集団の標本とみなして不偏分散を求める、VAR.S関数とVAR関数の使い方
- VAR.S　数値をもとに不偏分散を求める
- VAR.P　数値をもとに標本分散を求める
- VAR関数は、VAR.S関数の互換性関数です。

13

14

t分布

- 正規分布する母集団の平均と分散が未知で標本サイズが小さい場合に平均を推定する問題に利用

$$統計量T = \frac{標本平均(\overline{データ_n}) - \mu}{標本分散(S_n^2)/\sqrt{n}}$$

$$標本平均(\overline{データ_n}) = \frac{データ_1, データ_2, \cdots データ_n}{n}$$

$$標本分散(S_n^2) = \frac{1}{n-1}\sum_{i=1}^{n}(データ_i - \overline{データ_n})^2$$

$$これは f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}$$

$\nu =$ 自由度$(n-1)$と呼ばれる。

15

例題

- ある植物は3ヶ月後に10cmとなることが知られている。4つのサンプルの植物の平均は13cmでした。
  ・10cmと本質的に異なるかを知りたい=両側検定(短くなっている場合も考えられるため)
  ・10cmより伸びていることを知りたい=片側検定
- 一般的には片側確率より両側確率のほうが厳しい評価
- 両側検定
  - =T.DIST.2T(統計量,自由度)
  - =T.DIST.2T(3.18,3)
- 片側検定
  - =T.DIST(統計量,自由度,[累積分布／確率密度])
  - =T.DIST(-2.35,3,0)

16