第10章　テキストにおける推測分析

10.1　推定
10.2　仮説検定

09t4020a　　菊池裕紀

10.1　推定

- 推定とは？
  - 標本データの統計量(比率、平均、分散など)を用いて母集団の母数(母比率、母平均、母分散、)を推定すること。
  - 点推定

    標本データの統計量を母数とみなして推定する。
  - 区間推定

    標本データの統計量を用いて、母数が存在する区間を推定する。
    - 計算した統計量から何らかの確率分布に従うと仮定して推定する方法
    - シミュレーションにより推定する方法

区間推定

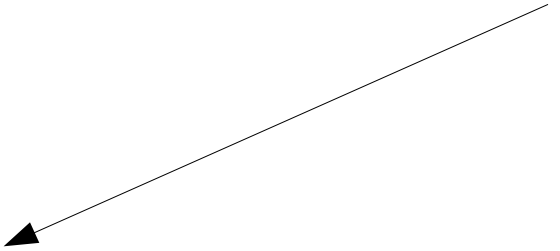
1.356	0.694	0.897	0.381	1.222	1.738
2.308	0.301	0.604	1.088	2.385	0.641
1.068	3.161	1.032	2.736	0.532	2.152
1.922	2.482	0.513	1.371	1.745	2.607
1.719	0.642	1.290	2.700	0.814	1.927

芥川龍之介の30編の作品について
1 文あたりに読点がどのくらい使用されているかを調べたデータ

平均の区間推定 (大標本)

- 標本サイズn が十分に大きい場合 (25 ≦ n)

平均μ
標準偏差σ
の母集団の標本平均 $\bar{x}$



平均μ、標準偏差 $\frac{\sigma}{\sqrt{n}}$ の正規分布に近づく

標準化 →

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

(標準正規分布$N(0,1)$に従う)

区間平均（大標本)

$$P\left(-1.96 \le 1.96\right) = P\left(-1.96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)$$

$$= P\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$= 0.95$$

100(1-α)

有意水準

※標準正規分布において、区間[-1.96,1.96]に含まれる確率は0.95の性質を利用

つまり、区間$\left[\bar{x}-1.96\frac{\sigma}{\sqrt{n}}, \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right]$に含まれる確率が0.95である事がわかる！

区間平均（小標本)

- スチューデントt分布
  - 標本サイズが小さく、母集団がどのゆな分布か分からない時、標本平均 $\bar{x}$ は自由度n-1のスチューデント分布(t分布)に従う。
  - 自由度30のt分布の区間[-2,2]に含まれる確率が0.95である。

$$\bar{x} - k\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + k\frac{\sigma}{\sqrt{n}}$$

※kは、自由度n-1のt分布のα/2の下・上側確率に対応する横軸の点の絶対値である。

10.2 仮説検定

- 仮説検定
  - 仮説を立て、それを統計的に立証する方法
  - 事前分布に基づいた統計的仮説検定のステップ

    1).仮説(帰無仮説)およびそれに対立する仮説をセットにしてたてる。

    2).検定すべき統計量に対応する検定統計量を求める。

    3).検定統計量が従う確率分布のもとで判断を下す。

統計的仮説検定

- 仮説に対して明確に判断(正しいor正しくない)する
  - 判断基準値として有意水準αを用いる！

- 仮説によって検定方法が変わる
  - 両側検定

    とにかく異なる事を立証する場合
  - 片側検定

    ある基準値よりも大きいか小さいかに関する事を立証する場合

母平均の検定

0.653	0.694	0.800	0.513	0.500	0.821	0.418	1.118	1.129
0.778	0.470	0.301	0.612	0.642	0.444	0.826	0.628	0.546
0.412	0.500	0.470	0.712	0.667	0.381	0.524	0.694	0.577
0.635	0.500	0.333	0.282	0.337				

芥川龍之介の32編の作品について
1 文あたりに読点がどのくらい使用されているかを調べたデータ

平均：0.591、標準偏差：0.206 ⟶ 全体作品における読点数より少ない

仮説検定を用いて分析する。

母平均の検定

- 全ての作品の母平均 μ_A
  - 帰無仮説 H_0 ： μ_A =0.591
  - 対立仮説 H_1 ： μ_A ≠0.591

    ここで、区間推定の標準化式zを用いて

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{0.591 - 1.499}{1.057 / \sqrt{32}} \fallingdotseq -4.859$$

  - zの値は標準正規分布に従うので……

    有意水準0.05とすると、α/2=0.025に対応する横軸の値は-1.96である。

平均の差の推定(大標本)

- 平均差の検定
  - 二つの標本が同じ母集団に属するかを検定

  - A：標準偏差(0.911)　母平均(μ_A)　母分散(σ_A^2)
  - B：標準偏差(0.661)　母平均(μ_B)　母分散(σ_B^2)

帰無仮説 H_0： $\mu_A = \mu_B$

対立仮説 H_1： $\mu_A \neq \mu_B$

平均差仮説検定(大標本)

- 大標本の場合の式

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sigma_{AB}} \quad , \quad \sigma_{AB} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

※　x：標本平均、n：標本サイズ

$$\sigma_{AB} \fallingdotseq \sqrt{\frac{0.661^2}{25} + \frac{0.206^2}{32}} = 0.137$$

$$z = \frac{0.911 - 0.591}{0.137} \fallingdotseq 2.33$$

平均差仮説検定(小標本)

- 母集団の分布が不明の時はt分布を用いる
  - 母分散が未知である両母平均の差のt検定統計量

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_{AB}} \quad , \quad s_{AB} \simeq \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

  - 自由度vのt分布に従う。
    - vが標本分散に等しい場合

$$v = n_A + n_B$$

    - vが標本分散に等しくない場合

$$v = \frac{\left(s_A^2/n_A + s_B^2/n_B\right)^2}{s_A^4/\left(n_A^2\left(n_A - 1\right)\right) + s_B^2/n_B^2\left(n_B - 1\right)}$$

割合の検定

- 二つの割合の差の検定
  - 仮説　H_0 ： $p_A = p_B$
  - 標本A,Bにおける、ある項目の割合がそれぞれ $\hat{p}_A, \hat{p}_B$ であり、標本サイズnが大きい時の検定等計量

$$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{p(1-p)(1/n_A + 1/n_B)}}$$

$$※ p \fallingdotseq \frac{\hat{p}_A n_A + \hat{p}_B n_B}{n_A + n_B}$$

割合の検定

		安倍	福田	zスコア	p値
日本	<名詞>	28	5	3.432	0.0003
国	<名詞>	32	7	3.3808	0.0004
立場	<名詞>	0	8	3.1476	0.0008
安心	<名詞>	2	12	3.0902	0.001
生活	<名詞>	1	10	3.0851	0.001
…		…	…	…	…
美しい	<形容詞>	9	1	2.2115	0.0135
文化	<名詞>	6	0	2.2052	0.0137
…		…	…	…	…
合計		2091	1692		

割合の検定

$$p = \frac{28+5}{2091+1692} = 0.0087$$

$$z = \frac{28/2091 - 5/1692}{\sqrt{0.0087 * (1-0.0087) * (1/2091 + 1/1691)}} \fallingdotseq 3.4$$