クレジット:
UTokyo Online Education 統計データ解析 I 2017 小池祐太

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

統計データ解析 (I) 第 12 回

小池祐太

2017 年 12 月 20 日

1. 検定
   - 仮説検定の考え方
   - 正規母集団に対する検定 (1 標本)
     - 平均の検定
     - 分散の検定
   - 正規母集団に対する検定 (2 標本)
     - 平均の差の検定
     - 分散の比の検定

2. 分散分析
   - 一元配置

検定

- **(統計的仮説) 検定**
  あるデータ/現象/母集団に対して仮定された命題 (仮説) の真偽を, そのデータの観測値に基づいて統計的に検証する方法
  - 例: 新しい薬の効能が古い薬よりも優れているといえるかということを, 薬の治験結果から検証したい
- 推定と大きく異なるのは, 母集団の分布に対して何らかの**仮説**を考えるところ
- 基本的な流れ
  1. 仮説を立てる
  2. 適当な統計量 (**検定統計量**と呼ばれる) に対して仮説が正しいときの標本分布を調べる
  3. 実際の検定統計量の値をデータから計算
  4. 計算された検定統計量の値が仮説に従う母集団から得られたと考えるに十分高い確率かどうかに基づいて仮説が正しいか否かを判断

検定: 用語

- **帰無仮説**: 検定統計量の分布を予想するために立てる仮説
  - 多くの場合「この仮説を捨てて無に帰したい」ことを期待して立てられるため, 「帰無」という言葉が使われる
- **帰無分布**: 帰無仮説が正しい場合の検定統計量の分布
  - 帰無仮説の下で検定統計量が取りうる値の範囲を予想するのに必要
- 帰無仮説を**棄却**する: 帰無仮説は誤っていると判断すること
- 帰無仮説を**受容**する: 帰無仮説を積極的に棄却することができないこと
- **棄却域**: 帰無仮説を棄却するために決める領域 (帰無仮説から予想される検定統計量の取りうる値の範囲外の領域)

検定: 用語 (続き)

- **第一種過誤**: 帰無仮説が正しいときに帰無仮説を棄却する誤り
- **第二種過誤**: 帰無仮説が誤っているときに帰無仮説を受容する誤り
- **有意水準**: 第一種過誤が起きる確率として許容する上限
- **サイズ**: 第一種過誤が起きる確率
- **検出力**: 第二種過誤が起きない確率
  - サイズを有意水準以下に抑えた上で, 可能な限り検出力を大きくするように棄却域をとるのが一般的な戦略
- **対立仮説**: 「帰無仮説が誤っているときに起こりうるシナリオ」として想定する仮説
  - 検出力のコントロールに必要
  - 慣習として, 帰無仮説を H_0, 対立仮説を H_1 で表すことが多い

検定

- 上の例では, 帰無仮説は,

$$H_0\text{: 新しい薬も古い薬も効能は同じ}$$

となり, 対立仮説としては例えば

$$H_1\text{: 新しい薬と古い薬の効能は異なる}$$

をとれる

- 対立仮説のとり方は別にある. 例えば, いまの場合, 新しい薬の効能は古い薬より良いことを期待しているので, 対立仮説として

$$H_1\text{: 新しい薬の方が古い薬より効能が高い}$$

をとるほうが妥当だと考えられる

検定

- 数式でまとめると, 棄却域は以下のようにして構成する
- 検定統計量を T, 有意水準を α とすれば, 上に述べたことから, 棄却域 R_α は, 帰無仮説が正しい場合に

$$P(T \text{ が棄却域 } R_\alpha \text{ に含まれる}) \leq \alpha \tag{1}$$

が成立するという制約の下で, 対立仮説が正しい場合に, 確率

$$P(T \text{ が棄却域 } R_\alpha \text{ に含まれる})$$

ができる限り大きくなるように定めるということ

- 帰無仮説のみによって棄却域が決まるのではなく, 対立仮説の立て方によって棄却域の形は変わり得ることに注意

検定

- 上記のように最初に有意水準を決めて棄却域を定める場合もあるが, データから計算された検定統計量の値に対して, その値が棄却域に含まれるような有意水準の最小値に基づいて検定を行う場合もある
- この値のことを p **値**という
- 数式で表すと,

$$\min\{\alpha \in (0,1) | \, T \text{ が } R_\alpha \text{に含まれる } \}$$

で与えられる (厳密には下限を考える)

- この場合, 検定の p 値が有意水準未満のときに帰無仮説を棄却することとなる

正規母集団に対する検定 (1 標本): 平均の検定

- 観測データ $X_1, X_2, \ldots, X_n$ が平均 μ, 分散 σ^2 の正規分布に従う独立同分布な確率変数列としてモデル化されている場合に, μ および分散 σ^2 に対する検定を行う方法を説明する
- まず, μ_0 を既知の定数として, 平均 μ が μ_0 であるか否かを検定する問題を考える
- 検定の用語を使って述べると, 帰無仮説を $\mu = \mu_0$, 対立仮説を $\mu \neq \mu_0$ とする検定を考える
- これはしばしば次の記号で表される:

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0. \tag{2}$$

正規母集団に対する検定 (1 標本): 平均の検定

- この仮説に対する検定は標本平均 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ が μ_0 からどの程度離れているかを検証することで行われる
- より具体的には, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ を不偏分散とし, 検定統計量として

$$t = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s}$$

  を考える
- 仮に帰無仮説 H_0 が正しいとすると, t は自由度 $n-1$ の t 分布に従う (配布資料 8.4.2 節参照)

正規母集団に対する検定 (1 標本): 平均の検定

- 従って, $\alpha \in (0,1)$ に対して, 自由度 $n-1$ の t 分布の $100(1-\alpha/2)\%$ 分位点を $t_{1-\alpha/2}(n-1)$ とすれば, H_0 の下では

$$P(|t| > t_{1-\alpha/2}(n-1)) = \alpha$$

  が成り立つ (配布資料 8.4.2 節参照)
- 以上より, 有意水準を α とする場合, 棄却域を

$$(-\infty, -t_{1-\alpha/2}(n-1)) \cup (t_{1-\alpha/2}(n-1), \infty)$$

  と設定すれば, 第一種過誤の上限が α となる

正規母集団に対する検定 (1 標本): 平均の検定

- 具体的な検定の手順としては, データから検定統計量 t の値を計算し,

$$|t| > t_{1-\alpha/2}(n-1)$$

  であった場合には帰無仮説を棄却する
- もしくは, 上で述べたように, 検定の p 値を計算して, p 値が α 未満の場合に帰無仮説を棄却するという手順をとってもよい
- いまの場合の検定の p 値は, $f(x)$ を自由度 $n-1$ の t 分布の確率密度関数として,

$$2\int_{|t|}^{\infty} f(x)dx \tag{3}$$

  によって与えられる

正規母集団に対する検定 (1 標本): 平均の検定

- $$|t| > t_{1-\alpha/2}(n-1) \Leftrightarrow \int_{-\infty}^{|t|} f(x)dx > 1-\alpha/2 \Leftrightarrow 2\int_{|t|}^{\infty} f(x)dx < \alpha$$

  が成り立つから, 検定統計量が棄却域に入ることと p 値が有意水準未満となることは同じ意味である
- (3) に現れる積分は関数 `pt()` のオプション `df` に自由度を, オプション `lower.tail` に `FALSE` を指定することで計算できるが, いまの場合は上の検定を実行するための関数 `t.test()` が p 値も計算してくれる

正規母集団に対する検定 (1 標本): 平均の検定

- なお, この検定のように, 帰無分布が t 分布となるような検定を t **検定**と呼ぶ
- 上の検定は **Student の** t **検定**と呼ばれることがある
- 実行例 `test-mean.r`

正規母集団に対する検定 (1 標本): 平均の検定

- 前節で述べたように, 対立仮説の立て方によって棄却域の形は変わりうる
- 例えば, 前節で述べた例に対応して, 観測データ $X_1, X_2, \ldots, X_n$ が新しい薬の効能を確認するためにその薬を n 人の被験者に投与した際の治験結果であったとする
  - 例えばその薬が睡眠薬であれば, データは睡眠時間の伸び具合に対応
- このとき, 母集団分布の平均 μ は新薬の「真の」, もしくは「平均的な」効能に対応するから, μ が古い薬の効能 μ_0 と比較して大きいと言えるのかどうかを考えるのが自然

正規母集団に対する検定 (1 標本): 平均の検定

- すなわち, 帰無仮説として $\mu = \mu_0$, 対立仮説として $\mu > \mu_0$ を設定した検定

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu > \mu_0.$$

  を考えるのが自然
- この場合, 帰無仮説は検定 (2) と同一なので, 検定統計量としても同一のもの t が利用出来る (帰無分布が計算可能であるため)
- 他方, 対立仮説の下では検定統計量 t の値が正の方向に大きくなると期待される
- 従って, 棄却域としては, c をある正の数として, "$t > c$" という形のものを考えるのが自然

正規母集団に対する検定 (1 標本): 平均の検定

- いま, $\alpha \in (0,1)$ に対して, 自由度 $n-1$ の t 分布の $100(1-\alpha)$%分位点を $t_{1-\alpha}(n-1)$ とすれば, H_0 の下で

$$P(t > t_{1-\alpha}(n-1)) = \alpha$$

  が成り立つ
- 以上より, 棄却域を

$$(t_{1-\alpha}(n-1), \infty)$$

  と設定すれば, 第一種過誤の上限が α となる

正規母集団に対する検定 (1 標本): 平均の検定

- 具体的な検定の手順としては, データから検定統計量 t の値を計算し,

$$t > t_{1-\alpha}(n-1)$$

であった場合には帰無仮説を棄却する

- もしくは, $f(x)$ を自由度 $n-1$ の t 分布の確率密度関数として, p 値

$$\int_t^\infty f(x)dx$$

が α 未満であった場合に帰無仮説を棄却するという手順をとっても同等

正規母集団に対する検定 (1 標本): 平均の検定

- 一般に, 棄却域がある定数 a によって (a, ∞) と書けるような検定を**右片側検定**と呼び, $(-\infty, a)$ と書けるような検定を**左片側検定**と呼ぶ. 右片側検定と左片側検定を合わせて**片側検定**と呼ぶ
- 一方で, 棄却域がある定数 $a < b$ によって $(-\infty, a) \cup (b, \infty)$ と書けるような検定を**両側検定**と呼ぶ
- いまの場合, 対立仮説を $H_1 : \mu \neq \mu_0$ にとった場合は両側検定となり, 対立仮説を $H_1 : \mu > \mu_0$ にとった場合は右片側検定となる

正規母集団に対する検定 (1 標本): 平均の検定

- 反対向きの対立仮説 $\mu < \mu_0$ を考えた検定

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu < \mu_0.$$

の場合は, 自由度 $n-1$ の t 分布の $100\alpha\%$分位点を $t_\alpha(n-1)$ として,

$$t < t_\alpha(n-1)$$

であった場合に帰無仮説を棄却すれば良い (左片側検定)

- これは, p 値

$$\int_{-\infty}^{t} f(x)dx$$

が α 未満であった場合に帰無仮説を棄却するということと同じである

- 実行例 `one-sided.r`

正規母集団に対する検定 (1 標本): 分散の検定

- 次に, σ_0^2 を既知の定数として, 分散 σ^2 が σ_0^2 であるか否かを検定する問題を考える
- 検定の用語を使って述べると, 帰無仮説を $\sigma^2 = \sigma_0^2$, 対立仮説を $\sigma^2 \neq \sigma_0^2$ とする検定

$$H_0 : \sigma^2 = \sigma_0^2 \qquad \text{vs} \qquad H_1 : \sigma^2 \neq \sigma_0^2$$

を考える

- $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ を不偏分散とする

正規母集団に対する検定 (1 標本): 分散の検定

- 仮に帰無仮説 H_0 が正しいとすると, 統計量

$$\chi^2 = (n-1)s^2/\sigma_0^2$$

は帰無仮説 H_0 の下で自由度 $n-1$ の χ^2 分布に従う (配布資料 8 章命題 8.2 参照)

- 従って, $\alpha \in (0,1)$ に対して, 自由度 $n-1$ の χ^2 分布の $100\alpha/2\%$分位点, $100(1-\alpha/2)\%$分位点をそれぞれ $\chi^2_{\alpha/2}(n-1)$, $\chi^2_{1-\alpha/2}(n-1)$ とすれば, H_0 の下では

$$P(\chi^2 < \chi^2_{\alpha/2}(n-1) \text{ または } \chi^2 > \chi^2_{1-\alpha/2}(n-1)) = \alpha$$

が成り立つ (配布資料 8.4.3 節参照)

正規母集団に対する検定 (1 標本): 分散の検定

- 以上より, 有意水準を α とする場合, 棄却域を

$$(-\infty, \chi^2_{\alpha/2}(n-1)) \cup (\chi^2_{1-\alpha/2}(n-1), \infty)$$

と設定すれば, 第一種過誤の上限が α となる

- 具体的な検定の手順としては, データから検定統計量 χ^2 の値を計算し,

$$\chi^2 < \chi^2_{\alpha/2}(n-1) \text{ または } \chi^2 > \chi^2_{1-\alpha/2}(n-1)$$

であった場合には帰無仮説を棄却する

正規母集団に対する検定 (1 標本): 分散の検定

- もしくは, この場合の p 値は, 自由度 $n-1$ の χ^2 分布の確率密度関数を $f(x)$ とすると

$$2\min\left\{\int_0^{\chi^2} f(x)dx, \int_{\chi^2}^{\infty} f(x)dx\right\}$$

で与えられるので, この値が α 未満の場合に帰無仮説を棄却するというのと同じ

正規母集団に対する検定 (1 標本): 分散の検定

- 対立仮説が片側の場合

$$H_0 : \sigma^2 = \sigma_0^2 \qquad \text{vs} \qquad H_1 : \sigma^2 > \sigma_0^2$$

を考えたときも, 前と同様の議論によって検定を構成できる

- すなわち, 自由度 $n-1$ の χ^2 分布の $100(1-\alpha)\%$分位点を $\chi^2_{1-\alpha}(n-1)$ として,

$$\chi^2 > \chi^2_{1-\alpha}(n-1)$$

であった場合に帰無仮説を棄却すれば良い

正規母集団に対する検定 (1 標本): 分散の検定

- 検定の p 値は

$$\int_{\chi^2}^{\infty} f(x)dx$$

で与えられるので, この値が α 未満の場合に帰無仮説を棄却するとしても同じ

- 左側対立仮説 $H_1 : \sigma^2 < \sigma_0^2$ の場合も同様
- なお, この検定のように, 帰無分布が χ^2 分布となるような検定を χ^2 **検定**と呼ぶ
- 実行例 `test-variance.r`

正規母集団に対する検定 (2 標本)

- 次に, 2 種類のデータに対する観測データがそれぞれ $X_1, X_2, \ldots, X_m$ および $Y_1, Y_2, \ldots, Y_n$ が与えられている状況で, 両者の平均や分散が一致するかどうかを検定する問題を考える
- 以下では次の 3 つの条件が満たされていると仮定する:
  1. $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$ は独立な確率変数列である.
  2. $X_1, X_2, \ldots, X_m$ は同分布であり, 平均 μ_1, 分散 σ_1^2 の正規分布に従う.
  3. $Y_1, Y_2, \ldots, Y_n$ は同分布であり, 平均 μ_2, 分散 σ_2^2 の正規分布に従う.

正規母集団に対する検定 (2 標本): 平均の差の検定

- まず, 2 種類のデータの平均が等しいか否かを検定する問題

$$H_0 : \mu_1 = \mu_2 \qquad \text{vs} \qquad H_1 : \mu_1 \neq \mu_2 \tag{4}$$

  を考える
- この問題は **Behrens-Fisher 問題**として知られており, 正確かつ適切な検定を導出することは難しいことが知られている
- そのため, 通常は以下で説明する **Welch の近似法** [1] と呼ばれる近似解が用いられる
  - χ^2 分布に従う独立確率変数列の一次結合の分布を, 1 つの χ^2 分布に従う確率変数の定数倍の分布で近似する方法
  - 平均と分散が元の確率変数と一致するように近似 (詳細は配布資料参照)

[1] **Satterthwaite の近似法**と呼ばれることもある.

正規母集団に対する検定 (2 標本): 平均の差の検定

- $X_1, \ldots, X_m$ の不偏分散を s_1^2, $Y_1, \ldots, Y_m$ の不偏分散を s_2^2 とする:

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2, \qquad s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

  - $\bar{X} - \bar{Y}, s_1^2, s_2^2$ は独立となることが知られている
  - また, 命題 8.2 より $(m-1)s_1^2/\sigma_1^2, (n-1)s_2^2/\sigma_2^2$ はそれぞれ自由度 $m-1, n-1$ の χ^2 分布に従う

正規母集団に対する検定 (2 標本): 平均の差の検定

- 確率変数 $s_1^2/m + s_2^2/n$ に Welch の近似法を適用すると, その分布は $c\chi_\nu^2$ の分布で近似できる. ただし, χ_ν^2 は自由度 ν の χ^2 分布に従う確率変数で,

$$c = \frac{(\sigma_1^2/m)^2}{m-1} + \frac{(\sigma_2^2/n)^2}{n-1}, \qquad \nu = \frac{(\sigma_1^2/m + \sigma_2^2/n)^2}{\frac{(\sigma_1^2/m)^2}{m-1} + \frac{(\sigma_2^2/n)^2}{n-1}}$$

- よって, 検定統計量

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_1^2/m + s_2^2/n}}$$

  を考えると, t の分布は $(\bar{X} - \bar{Y})/\sqrt{c\chi_\nu^2}$ で近似できる
- さらに, $\bar{X} - \bar{Y}, s_1^2/m + s_2^2/n$ は独立であるから, $\bar{X} - \bar{Y}, c\chi_\nu^2$ も独立とすべきである

正規母集団に対する検定 (2 標本): 平均の差の検定

- 命題 8.1 より $((\bar{X} - \bar{Y}) - (\mu_1 - \mu_2))/\sqrt{\sigma_1^2/m + \sigma_2^2/n}$ は標準正規分布に従うから, H_0 の下で

$$\frac{\bar{X} - \bar{Y}}{\sqrt{c\chi_\nu^2}} \left(= \frac{(\bar{X} - \bar{Y})/\sqrt{\sigma_1^2/m + \sigma_2^2/n}}{\sqrt{\frac{c}{\sigma_1^2/m + \sigma_2^2/n}\chi_\nu^2}} \right)$$

  は自由度 ν の t 分布に従う ($\frac{c}{\sigma_1^2/m + \sigma_2^2/n} = \nu^{-1}$ に注意)

- 従って, 元々の検定統計量 t の帰無分布は自由度 ν の t 分布で近似できることになる
- ν は未知の分散 σ_1^2, σ_2^2 を含むので, 実際の応用ではこれらを不偏推定量 s_1^2, s_2^2 で代用して, 次式で与えられる自由度 $\hat{\nu}$ を用いる:

$$\hat{\nu} = \frac{(s_1^2/m + s_2^2/n)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}.$$

正規母集団に対する検定 (2 標本): 平均の差の検定

- 具体的な検定の手順は以下の通り
  - 有意水準を $\alpha \in (0,1)$ とする場合, 自由度 $\hat{\nu}$ の t 分布の $100(1-\alpha/2)$%分位点を $t_{1-\alpha/2}(\hat{\nu})$ として, 棄却域を

$$(-\infty, -t_{1-\alpha/2}(\hat{\nu})) \cup (t_{1-\alpha/2}(\hat{\nu}), \infty)$$

    と設定する
  - 従って, データから検定統計量 t の値を計算し,

$$|t| > t_{1-\alpha/2}(\hat{\nu})$$

    であった場合には帰無仮説を棄却する
- この検定は **Welch の t 検定**と呼ばれることがある
  - p 値の計算方法や片側対立仮説の場合への対応方法は前と類推の議論となるため省略
- 実行例 `test-welch.r`, `kion-difference.r`

平均の差の検定 (対応がある場合)

- 2 種類のデータを考える場合, 2 つのデータ間に自然な対応を考えることができることがある
  - 例えば, 2 種類の薬の効能を比較するために, n 人の被験者にそれぞれの薬を投与したとする
  - このとき, 各 $i = 1, \ldots, n$ について, i 番目の被験者にそれぞれの薬を投与した場合の治験結果を X_i, Y_i とした場合, X_i と Y_i には「同一の被験者に対する治験結果」という意味で対応がある
- このような場合, 仮説検定 (4) の代わりに,「対応がある観測値の差の平均が 0 か否か」という仮説検定を考えることができる
  - すなわち, $Z_i = X_i - Y_i$ $(i = 1, \ldots, n)$ として, $Z_1, \ldots, Z_n$ たちの平均が 0 か否かを上で述べた方法で検定すれば良い
- 実行例 `test-paired.r`

正規母集団に対する検定 (2 標本): 分散の比の検定

- 2 種類のデータの分散が等しいか否かを検定する問題

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad \text{vs} \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

  を考える
- $X_1, \ldots, X_m$ の不偏分散を s_1^2, $Y_1, \ldots, Y_m$ の不偏分散を s_2^2 とする
- s_1^2, s_2^2 は独立であり, また命題 8.2 より $(m-1)s_1^2/\sigma_1^2, (n-1)s_2^2/\sigma_2^2$ はそれぞれ自由度 $m-1, n-1$ の χ^2 分布に従う
- 従って, 検定統計量として

$$F = s_1^2/s_2^2$$

  を考えると, H_0 の下で F は自由度 $m-1, n-1$ の F 分布に従う (6.3.6 節参照)

正規母集団に対する検定 (2 標本): 分散の比の検定

- よって, $\alpha \in (0,1)$ に対して, 自由度 $m-1, n-1$ の F 分布の $100\alpha/2$%分位点, $100(1-\alpha/2)$%分位点をそれぞれ $F_{\alpha/2}(m-1,n-1)$, $F_{1-\alpha/2}(m-1,n-1)$ とすれば, H_0 の下では

$$P(F < F_{\alpha/2}(m-1,n-1) \text{ または } F > F_{1-\alpha/2}(m-1,n-1)) = \alpha$$

  が成り立つ
- 以上より, 有意水準を α とする場合, 棄却域を

$$(-\infty, F_{\alpha/2}(m-1,n-1)) \cup (F_{1-\alpha/2}(m-1,n-1), \infty)$$

  と設定すれば, 第一種過誤の上限が α となる

正規母集団に対する検定 (2 標本): 分散の比の検定

- 具体的な検定の手順としては, データから検定統計量 F の値を計算し,

$$F < F_{\alpha/2}(m-1, n-1) \text{ または } F > F_{1-\alpha/2}(m-1, n-1)$$

  であった場合には帰無仮説を棄却する
- p 値の計算方法や片側対立仮説の場合への対応方法は前と類推の議論となるため省略
- 実行例 `test-ratio.r`, `kion-ratio.r`

分散分析

- 上で述べた平均の差の検定は, 2 つのグループ間で平均の差があるか否かを検定する方法であった
- **分散分析**とは, 大雑把にいうと, 2 つ以上のグループ間で平均の差があるか否かを検定する方法
  - 例えば, ある小売店について,「売上高は月によって差があるか」という仮説を検定したり, また, ある銘柄の株価について「収益率は曜日によって差があるか」という仮説を検定するのに分散分析は有用

分散分析

- 分散分析の基本的な考え方は, データの変動からグループ間での変動と観測誤差のみに起因する変動を抽出し, 両者を比較すること
  - もしグループ間で平均に差がなければ, グループ間での変動は観測誤差のみに起因する変動と自由度を除いて本質的な差がないはず
  - 逆にグループ間で平均に差があれば, 前者はその分だけ変動が増して後者より大きくなるはずなので, 両者の比較によって目的の検定が実行できる
- 従って, 分散分析は「分散の分析」というよりむしろ「データの変動の分析」であるといえる

一元配置

- まず, グループ分けが 1 種類の場合を考え, p 個のグループ $A_1, A_2, \ldots, A_p$ があるとする
- 統計学では, グループ分けのことを**因子**と呼び, 因子内の各グループのことを**水準**と呼ぶことが多いため, 以下これらの用語を用いる
- 各 $i = 1, 2, \ldots, p$ について n_i 個の観測データ $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ が与えられている状況を考える (例えば, $A_1, A_2, \ldots, A_p$ が月に対応し, $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ が i 月の各日における売上高に対応していると考えれば良い)

一元配置

- 観測データは以下のモデルに従うと仮定する:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (i = 1, \ldots, p;\ j = 1, \ldots, n_i). \tag{5}$$

  - μ_i は定数であり, 水準 A_i における観測データの平均値を表す
  - ε_{ij} は確率変数であり, $\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \ldots, \varepsilon_{p1}, \ldots, \varepsilon_{pn_p}$ は独立同分布で平均 0, 分散 σ^2 の正規分布に従うと仮定する
- 水準 $A_1, A_2, \ldots, A_p$ の間の平均値に差があるか否かを検定する問題は, 以下のように定式化できる:

$$H_0 : \mu_1 = \cdots = \mu_p \qquad \text{vs} \qquad H_1 : \text{ある } i, j \text{ に対して} \mu_i \neq \mu_j.$$

一元配置

- 冒頭で述べたように, 分散分析ではデータの変動から因子間での変動と観測誤差のみに起因する変動を抽出し, 両者を比較することで検定を構成する
- まず, データ全体の標本平均 $\bar{Y}_{\cdot\cdot}$ および水準 A_i における標本平均 $\bar{Y}_{i\cdot}$ を以下で定義する:

$$\bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}, \qquad \bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad (i = 1, \ldots, p).$$

ただし, $n := \sum_{i=1}^{p} n_i$ は全サンプル数を表す

一元配置

- 次に, 各水準内でのデータの変動 (の合計)SS_W, 水準間でのデータの変動 SS_B を以下で定義する:

$$SS_W = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2,$$

$$SS_B = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{p} n_i (Y_{i\cdot} - \bar{Y}_{\cdot\cdot})^2.$$

  - SS_W を**級内変動**, SS_B を**級間変動**と呼ぶ
  - いまの設定では, 級内変動 SS_W は観測誤差にみに起因して生じる

一元配置

- 仮に帰無仮説 H_0 が正しければ, 水準内でのデータの変動・水準間でのデータの変動ともに観測誤差のみが原因で生じるはずなので, 自由度を除けば本質的な違いはないはずである
- 逆に, 対立仮説 H_1 が正しければ, 水準間でのデータの変動は観測誤差のみならず, 水準間での平均値 $\mu_1, \ldots, \mu_p$ の異質性にも影響されるはずなので, SS_B は SS_W より本質的に大きくなるはずである
- 数学的には, 帰無仮説の下で, $SS_W/(n-p)$, $SS_B/(p-1)$ はともに σ^2 の不偏推定量となることが示せる

一元配置

- 従って, 検定統計量として

$$F = \frac{SS_B/(p-1)}{SS_W/(n-p)}$$

  を考えるのが自然である
- 対立仮説の下では F は大きな値をとるはずなので, この検定は右片側検定となる
- 帰無仮説の下で次の事実が成り立つことが知られている: SS_B, SS_W は独立であり, SS_B は自由度 $p-1$ の χ^2 分布に従い, SS_W は自由度 $n-p$ の χ^2 分布に従う
- 従って, 帰無仮説の下で F は自由度 $p-1, n-p$ の F 分布に従う (6.3.6 節参照)

一元配置

- よって, $\alpha \in (0,1)$ に対して, 自由度 $p-1, n-p$ の F 分布の $100(1-\alpha)$%分位点を $F_{1-\alpha}(p-1, n-p)$ とすれば, H_0 の下では

$$P(F > F_{1-\alpha}(p-1, n-p)) = \alpha$$

  が成り立つ
- 以上より, 有意水準を α とする場合, 棄却域を

$$(F_{1-\alpha}(p-1, n-p), \infty)$$

  と設定すれば, 第一種過誤の上限が α となる

一元配置

- 具体的な検定の手順としては, データから検定統計量 F の値を計算し,

$$F > F_{1-\alpha}(p-1, n-p)$$

であった場合には帰無仮説を棄却する

- もしくは, $f(x)$ を自由度 $p-1, n-p$ の F 分布の確率密度関数として, p 値

$$\int_F^\infty f(x)dx$$

が α 未満であった場合に帰無仮説を棄却するという手順をとっても同等

分散分析表 (一元配置の場合)

	自由度	平方和	平均平方和	F 値	p 値
級間	$p-1$	SS_B	$SS_B/(p-1)$	F	$\int_F^\infty f(x)dx$
級内	$n-p$	SS_W	$SS_W/(n-p)$		

一元配置

- モデル (5) では各水準の効果をその水準における平均値で表していたが, 因子 A 全体の平均効果を μ で表して, 平均 μ を基準とした各水準 A_i の相対的な効果 α_i で表すことも可能
- すなわち,

$$\mu_i = \mu + \alpha_i, \qquad \mu = \frac{1}{n}\sum_{i=1}^{p} n_i \mu_i$$

とする

一元配置

- このとき,

$$\sum_{i=1}^{p} n_i \alpha_i = 0$$

であるから, 帰無仮説 H_0 は

$$\alpha_1 = \cdots = \alpha_p = 0$$

と同等

- R には分散分析を実行するための関数 `aov()` が用意されている
- 実行例 `anova-oneway.r`