# 第1回 : 多変量解析の概要と線形回帰の復習

Code

# 1. 多変量解析の概要

**多変量解析**（multivariate analysis）という言葉は、複数の変数からなるデータを分析するために考案された様々な手法の総称です。続・初級統計学では、この「多変量解析」について詳しく学んでいきます。

## 1.1 データ

一般にデータは、各行に複数の変数を観測した結果を表にまとめた形で表されます。データの一例として、講義資料の data ディレクトリにある salary.csv という名前のcsvファイルを開けてみてください。このデータはある会社の社員に関する情報を記録したデータです。

Rに読み込んでから確認することもできます。Rにデータを読み込むときは

1. 作業ディレクトリの指定
2. 以下のスクリプトを実行

という流れをおこないます。作業ディレクトリの指定は、RStudioの場合、 Session -> Set Working Directory -> Choose Directory と進むことで可能です。今回は、 lecture01 ディレクトリを作業ディレクトリに指定してください。ここまで出来たら、以下のスクリプトを実行してみましょう。

Hide

```
# データが表の形で表される様子
dat <- read.csv(file = "./data/salary.csv",
                fileEncoding = "utf-8")
head(x = dat, n = 5)    # 先頭5行を確認する。
```

| | 月給 <int> | 勤続年数 <int> | 仕事の達成度 <int> | 欠勤日数 <int> | 特殊免許の有無 <int> |
|---|---|---|---|---|---|
| 1 | 245803 | 5 | 144 | 6 | 0 |
| 2 | 241428 | 1 | 234 | 5 | 1 |
| 3 | 272966 | 6 | 240 | 5 | 1 |
| 4 | 326961 | 19 | 207 | 4 | 1 |
| 5 | 240608 | 1 | 189 | 5 | 1 |

5 rows

**Remark** : エラーが出力され、 ファイル './data/salary.csv' を開くことができません: No such file or directory という警告メッセージが出力されることがあります。この場合、

- ファイル名に綴りのミスがある。
- ディレクトリの変更に失敗している。

などが主な失敗にあげられます。ファイル名に綴りのミスが見当たらない場合、作業ディレクトリを getwd() 関数で出力し、 lecture01 になっているか確認してみてください。■

データの各行を**標本点**（sample point）といい、標本点の個数を**標本サイズ**（sample size）といいます。また各列を **変数** （variable）といい、変数の個数を**データの次元**（dimension）といいます。

**問題** : 先ほどRに読み込んだデータ salary.csv について、標本サイズと次元を答えてください。また、変数の名前をすべて列挙してください。Hint : str 関数を用いると便利です。

**解答** :
以下のスクリプトを実行してみましょう。

Hide

```
str(dat)
```

```
'data.frame':   100 obs. of  5 variables:
 $ 月給           : int  245803 241428 272966 326961 240608 261362 275045 236078 258850 280603 ...
 $ 勤続年数       : int  5 1 6 19 1 5 6 4 3 9 ...
 $ 仕事の達成度   : int  144 234 240 207 189 212 224 173 263 214 ...
 $ 欠勤日数       : int  6 5 5 4 5 5 4 7 6 4 ...
 $ 特殊免許の有無: int  0 1 1 1 1 1 1 0 1 0 ...
```

この出力から、標本サイズは 100 、次元は 5 、変数の名前は
月給 、 勤続年数 、 仕事の達成度 、 欠勤日数 、 特殊免許の有無 だとわかります。■

# 1.2 多変量解析の様々な課題と手法

**問題** : salary.csv は、ある会社の社員に関する 月給 、 勤続年数 、 仕事の達成度 、 欠勤日数 、 特殊免許の有無 の情報が記録されたデータです。このデータから分析できる課題を考えてみましょう。（できれば複数考えてみてください。）

**解答** : 例を4つほど掲げます。

1. 社員の 月給 の傾向が 特殊免許の有無 によって異なるかを検討する。
2. 社員の 月給 の傾向を 勤続年数 、 仕事の達成度 、 欠勤日数 、 特殊免許の有無 の4変数から説明できないか検討する。
3. 勤続年数 、 仕事の達成度 、 欠勤日数 、 特殊免許の有無 の4変数を用いて、社員の「熟練度」を1つの数値で表すような指標を作れないか検討する。
4. 仕事の達成度 、 欠勤日数 、 特殊免許の有無 の3変数を用いて、似たような勤務態度の社員をグルーピングできないか検討する。

以上は一例です。他にも様々な課題を考えることができます。■

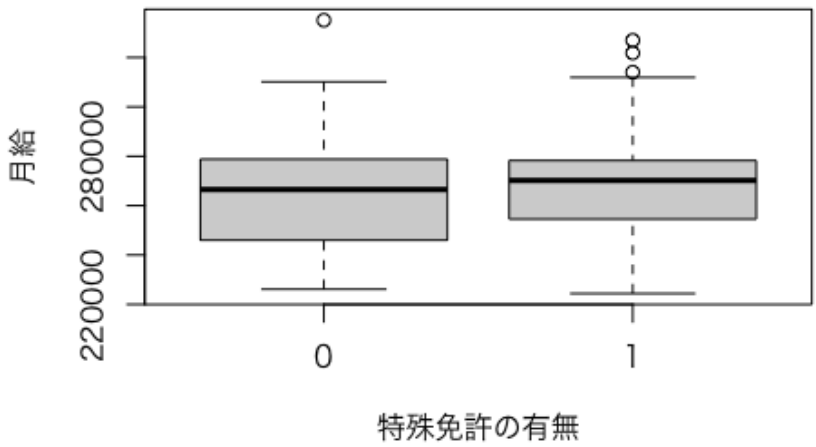
解答 (1) は、初級統計学で習う手法によって検討できます。例えば、

1. 横軸を 特殊免許の有無 、縦軸を 月給 とした「箱ひげ図」をかいて両群の月給を比較する。
2. 統計的仮説検定、今回の場合は母平均の差の検定を行う。

という流れが考えられます。箱ひげ図から、両群の月給の分布には、あまり差がないことが見てとれます。また、母平均の差の検定からも帰無仮説は棄却されず、両群の月給の母平均に差があるとは言えないという結論になります。

Hide

```
# 箱ひげ図
par(family = "ヒラギノ角ゴシック W3")    # Macユーザーのみ
boxplot(月給 ~ 特殊免許の有無, data = dat)
```




Hide

```
# 母平均の差の検定（両側検定、有意水準5%に設定した。）
t.test(月給 ~ 特殊免許の有無, data = dat,
         alternative = "two.sided", conf.level = 0.95)
```

```
    Welch Two Sample t-test

data:  月給 by 特殊免許の有無
t = -0.87961, df = 97.957, p-value = 0.3812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12553.174   4842.607
sample estimates:
mean in group 0 mean in group 1
       265917.4        269772.7
```

一方、(2), (3), (4)のように3つ以上の変数を対象とするような検討課題に対しては、初級統計学では触れてきませんでした。このような課題に対して有効な手法が「多変量解析」です。

さて、多変量解析と一口に言っても、想定される課題は様々あることがわかります。解答の(2)は予測や説明を目的とした課題であるのに対し、解答(3)や解答の(4)は変数間の共通性や標本点間の類似性の抽出といったデータの構造の理解を目指しています。また、各課題に対応する多変量解析の手法も一つとは限りません。様々な観点から、たくさんの手法が提案されています。なお続・初級統計学では、以下に掲げる多変量解析の手法に絞って解説していきます。

- 線形回帰
- 累乗モデル
- ロジスティック回帰
- 判別分析
- 主成分分析
- 因子分析
- k-means法
- Ward法
- 決定木とアンサンブル学習
- サポートベクトルマシン
- ニューラルネットワークと深層学習
- 多次元尺度構成法
- コレスポンデンス分析
- 共分散構造分析（構造方程式モデリング）

**Remark** : このように、多変量解析は多様性に富んだ一大分野です。多変量解析の勉強に挫折する方の多くは、この多様性に圧倒されているようにも思います。初学者の方は、まず様々な多変量解析の手法について、その概要をおさえることを目指してみてください。特に、

- どのような課題に対する手法なのか。
- どのような結果を計算する手法なのか。
- どのような仕組みになっているのか。

この3点をよく把握することをお勧めします。

# 2. 線形回帰の復習

初級統計学では「回帰分析」、正確には線形回帰とよばれる手法について勉強しました。その復習から始めていきましょう。

# 2.1 線形回帰の概要

## A. 課題設定

データに $D$ 個の変数 $x_1, \cdots, x_D$ と $y$ が記録されているとき、変数 $y$ を変数 $x_1, \cdots, x_D$ の1次式

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D + \text{誤差}$$

で表現することで説明・予測できると仮定します。例えば、社員の 月給 の傾向を
勤続年数 、 仕事の達成度 、 欠勤日数 、 特殊免許の有無 の4変数から説明できないか検討したい場合、$y$ を 月給 、$x_1, \cdots, x_4$ を 勤続年数 、…、 特殊免許の有無 と置き換え

| id | $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|---|
| 1 | 245803 | 5 | 144 | 6 | 0 |
| 2 | 241428 | 1 | 234 | 5 | 1 |
| 3 | 272966 | 6 | 240 | 5 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

この場合、

$$\text{月給} = \beta_0 + \beta_1 \times \text{勤続年数} + \beta_2 \times \text{仕事の達成度} + \beta_3 \times \text{欠勤日数} + \beta_4 \times \text{特殊免許の有無} + \text{誤差}$$

という式を考えています。$x_1, \cdots, x_D$ を**説明変数**（explanatory variable）、$y$ を**目的変数**（objective variable） といいます。

## B. 出来ること

$\beta_0, \beta_1 \cdots, \beta_D$ を**偏回帰係数**（coefficient）、特に $\beta_0$ を**切片**（intercept）といいます。これらの係数を推定することで

- 説明変数 $x_1$ だけが $1$ 異なり、他の説明変数 $x_2, \cdots, x_D$ が同じ場合、目的変数 $y$ の値の差はどの程度か？
- 各説明変数の値が決まっているとき、目的変数の値はいくらと予測できるか？

を推定することができます。また、偏回帰係数の区間推定や統計的仮説検定を行うことができます。これについてはデモで例を交えながら解説します。

## C. 偏回帰係数の推定の仕組み

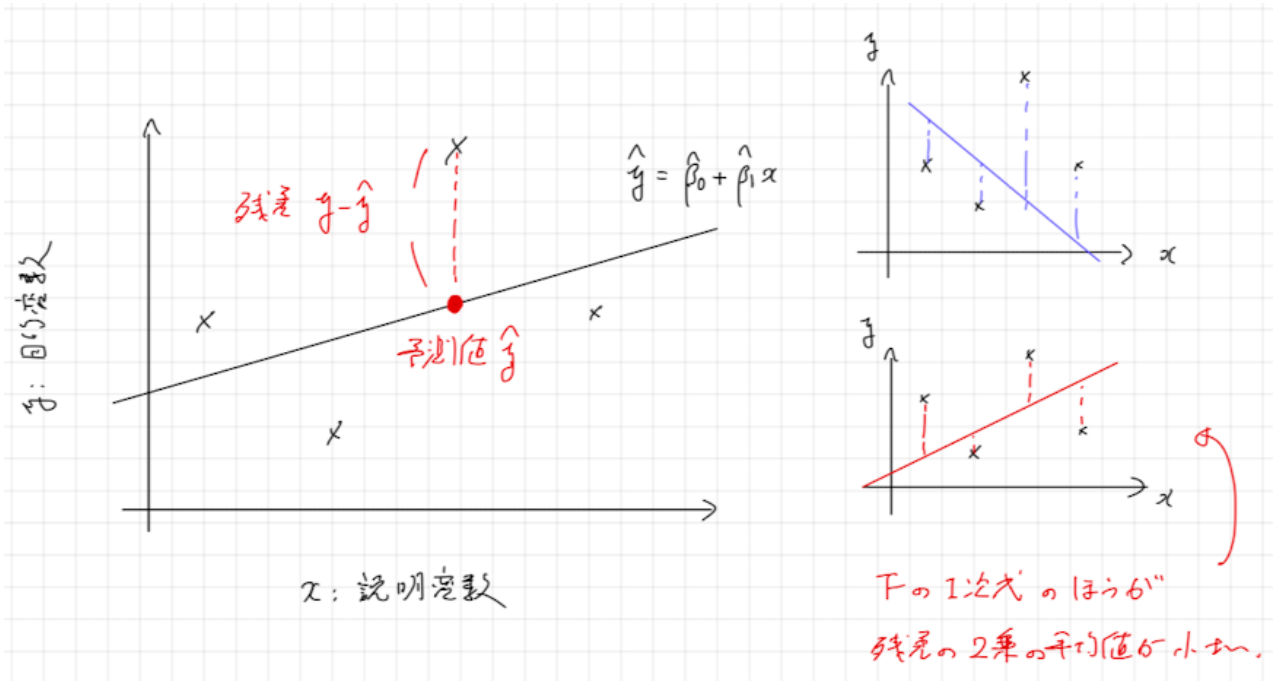
偏回帰係数の推定値 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_D$ は、データ全体で

- 偏回帰係数から決まる各標本点の予測値 $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x_1 + \cdots + \hat{\beta_D} x_D$
- 実測値 $y$

の差の2乗（**残差**といいます） $(y - \hat{y})^2$ の平均値

$$l(\beta_0, \beta_1, \cdots, \beta_D) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

が最小になるように決めます。これを**最小2乗法**（least squares method）といいます。線形回帰とは、目的変数と説明変数の関係に1次式を仮定したとき、これを最小2乗法で推定する手法のことです。




**Remark** : 2乗は符号を打ち消すためについています。それなら、絶対値 $|y_i - \hat{y_i}|$ でも良いではないかと思う方がいるかもしれません。実際に、$l^1$-損失といってこちらを使うこともあるのですが、今回は詳しく触れないことにします。■

# 2.2 線形回帰のデモ

salary.csv は、ある会社の社員に関する 月給 、 勤続年数 、 仕事の達成度 、 欠勤日数 、 特殊免許の有無 の情報が記録されたデータです。今回、社員の 月給 の傾向を 勤続年数 、 仕事の達成度 、 欠勤日数 、 特殊免許の有無 の4変数から説明できないか検討していきます。今回は、月給 $y$ を

$$\text{月給} = \beta_0 + \beta_1 \times \text{勤続年数} + \beta_2 \times \text{仕事の達成度} + \beta_3 \times \text{欠勤日数} + \beta_4 \times \text{特殊免許の有無} + \text{誤差}$$

で説明するモデル、つまり線形回帰を考えます。

## step 1. 線形性の確認

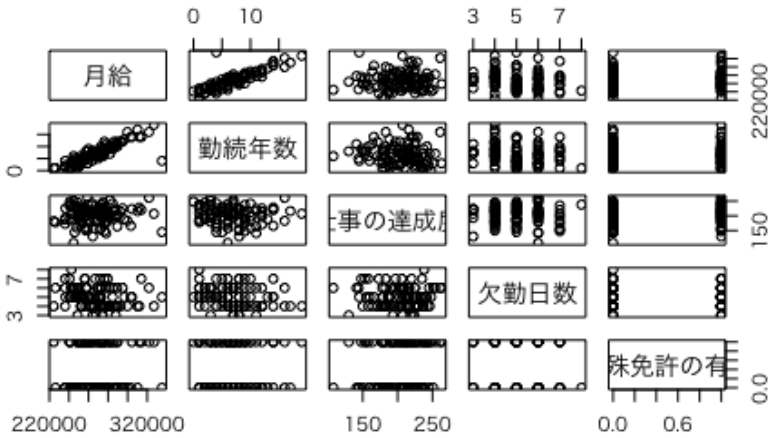
まず、以下のような1次式（線形性）を仮定することが妥当なのかを検討しておきましょう。

$$\text{月給} = \beta_0 + \beta_1 \times \text{勤続年数} + \beta_2 \times \text{仕事の達成度} + \beta_3 \times \text{欠勤日数} + \beta_4 \times \text{特殊免許の有無} + \text{誤差}$$

これは散布図をかくことで検討することができます。

Hide

```
# 散布図
par(family = "ヒラギノ角ゴシック W3")    # Macユーザーのみ
plot(dat)
```




**問題** : 散布図から「線形性」を仮定しても問題ないかを検討してください。

**解答** : 検討の一例を紹介します。散布図から目的変数と説明変数の間には、それぞれ線形の関係がみられることが確認できます。これより「線形性」を仮定してもよいと考えられます。■

# B. 偏回帰係数の推定

切片および偏回帰係数 $\beta_0, \beta_1, \cdots, \beta_4$ の推定値を求めます。R言語では、 lm 関数を用いて線形回帰を計算できます。以下のスクリプトを確認してください。

Hide

```
# 線形回帰はlm関数で計算できます。
result <- lm(formula = 月給 ~ ., data = dat)
summary(result)
```

```
Call:
lm(formula = 月給 ~ ., data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-15413  -3884     -5   3004  88154

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    213082.90    9511.38  22.403  < 2e-16 ***
勤続年数         4761.56     271.51  17.538  < 2e-16 ***
仕事の達成度       93.07      35.25   2.640  0.00969 **
欠勤日数          262.71     975.75   0.269  0.78833
特殊免許の有無   3977.66    2150.76   1.849  0.06751 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10630 on 95 degrees of freedom
Multiple R-squared:  0.7746,    Adjusted R-squared:  0.7651
F-statistic: 81.61 on 4 and 95 DF,  p-value: < 2.2e-16
```

偏回帰係数の推定値は Coefficients の Estimate に書かれています。 Intercept は切片という意味です。計算の結果、$\hat{\beta_0} = 213083, \hat{\beta_1} = 4762, \hat{\beta_2} = 93, \hat{\beta_3} = 263, \hat{\beta_4} = 3978$ だとわかります。具体的に1次式で表すと、

$$\text{月給} = 213083 + 4762 \times \text{勤続年数} + 93 \times \text{仕事の達成度} + 263 \times \text{欠勤日数} + 3978 \times \text{特殊免許の有無} +$$

が得られたことになります。

**問題** : 以下の空欄（ア），（イ）にあてはまる値を答えてください。

- 勤続年数 ・ 仕事の達成度 ・ 特殊免許の有無 が同じ二人の社員について、 欠勤日数 が1日多い社員は、そうでない社員に比べて 月給 が（ア）円だけ高い傾向にあると推定される。
- 勤続年数 2年、 仕事の達成度 200、 欠勤日数 3日、 特殊免許 を持っていない人の 月給 は（イ）円と予測できる。

**解答** : （ア）の値は 263、（イ）の値は

$$213083 + 4762 \times 2 + 93 \times 200 + 263 \times 3 + 3978 \times 0 = 241996$$

とわかります。■

とこで、欠勤日数の偏回帰係数の推定値 $\hat{\beta}_3$ は 263 と正の値になっていることに違和感を覚えた方はいるでしょうか。よく読むと、欠勤日数の多い人の方が月給が高い傾向にあるという計算結果になっています。これは直感と異なる結果でしょう。直感や事前知識と異なる結果に気づくことは、

- データのとり方に不備があったのではないか？
- 説明変数に不足があり、欠勤日数に相関のある別の変数の影響が $\hat{\beta}_3$ の値に現れているのではないか？

など、データを理解するきっかけになります。「そうか！欠勤日数が多いほど月給が多い傾向にあるのか！これは発見だ！」とモデルを過信せず、結果から読み取れることに納得いくまで、その結果が出る理由をよく考察することが、多変量解析では大切です。

# C. 偏回帰係数の区間推定・統計的仮説検定

データ分析を担当しているSさんは、特殊免許を持っている人は持っていない人に比べて、月給を 3000円より多く貰う傾向にあるのではないかと考えていました。この仮説を検討するための統計的仮説検定を行ってみましょう。有意水準は $5\%$ とします。

帰無仮説 $H_0 : w_d = w$ に対して、検定統計量を (Estimate-w)/Std.Error とします。帰無仮説が正しいとき、この検定統計量は自由度 $n - (D + 1)$ の $t$ 分布に従います。この事実を用いて、偏回帰係数の統計的仮説検定や区間推定を行うことができます。

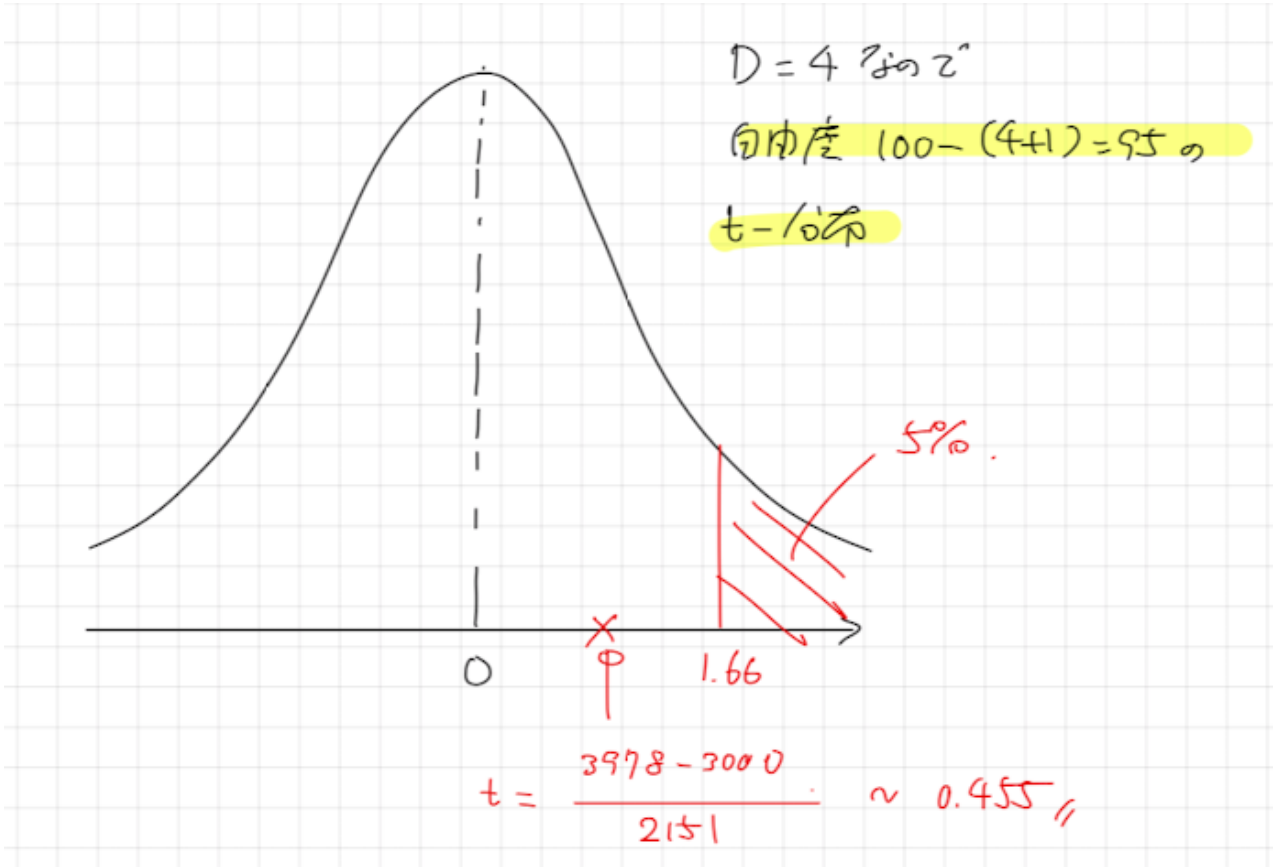
今回、帰無仮説と対立仮説をそれぞれ

$$H_0 : \beta_3 = 3000 \text{ v.s. } H_1 : \beta_3 > 3000$$

とします。このとき、検定統計量の値は

$$t = \frac{3978 - 3000}{2151} \sim 0.455$$

です。片側検定であることに注意すると、有意水準 $5\%$ のとき棄却域は$t \geq 1.66$ （R言語を用いて qt(p=0.95, df=100-5) と計算できます）です。




すなわち、帰無仮説は棄却されません。すなわち「特殊免許を持っている人は持っていない人に比べて、月給を3000円より多く貰う傾向にあるとは言えない」という結果になります。■

**Remark** : もし対立仮説が $H_1 : w_3 \neq 3000$ の場合、両側検定であることに注意すると、棄却域は $|t| \geq 1.985$ になります。