12. 2 つの変数の関係を理解する

honocat

2025-12-26

データの入手。

```
myd <- read_csv('data/fake_bivariate.csv')
names(myd)
```

```
[1] "id"       "female"   "support"  "height"   "faheight"
```

```
glimpse(myd)
```

```
Rows: 1,000
Columns: 5
$ id       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
$ female   <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1~
$ support  <dbl> 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0~
$ height   <dbl> 174.5, 171.3, 174.2, 173.5, 169.7, 157.0, 158.5, 161.5, 164.1~
$ faheight <dbl> 179.0, 166.4, 169.7, 171.5, 156.5, 163.1, 170.5, 173.8, 180.7~
```

質的変数同士の関係を調べる

質的変数

myd には女性を表す female 変数と、内閣に対する指示を表す support 変数がある。これらの変数では、数字自体には特に意味がない。このような変数は、質的変数（qualitative variables）と呼ばれる。また、female のようにある特性を備えているかどうかを 0 と 1 で表現する変数を**ダミー変数**（dummy variable）と呼ぶ。

```
table(myd$female)
```

```
  0   1
500 500
```

0 と 1 だと何を表しているかわかりにくい。質的変数を **factor** 型にする。

```
myd <- myd |>
  mutate(female = factor(female,
                         levels = c(0, 1),
                         labels = c('male', 'female')))
table(myd$female)
```

```
  male female
   500    500
```

内閣への指示も変換する。

```
myd <- myd |>
  mutate(support = factor(support,
                          levels = c(0, 1),
                          labels = c('dis', 'sup')))
table(myd$support)
```

```
dis sup
450 550
```

クロス表

上で見た 2 つの質的変数同士には、なにか関係があるのだろうか。これを確かめるために、2 つの質的変数の関係を表にしてみよう。複数の変数を使った表を**クロス表（cross table, contingency table）**という。

```
table(myd$female, myd$support)
```

```
        dis sup
  male   200 300
  female 250 250
```

```
# with(myd, table(female, support))
```

この表に名前をつけて保存し、後で使えるようにする。

```
tbl_fem_sup <- with(myd, table(female, support))
```

男性で内閣不支持が 200 人、男性で指示するのが 300 人、女性では不支持も支持も 250 人ずついることがわかる。

合計値も追加する。合計値は、データにとっては周辺的な情報であるし、物理的に見ても表の周辺に記載されるので、**周辺度数（margins）**と呼ばれる。

```
addmargins(tbl_fem_sup) # margin 引数で行列を指定
```

```
        support
female    dis  sup  Sum
  male    200  300  500
  female  250  250  500
  Sum     450  550 1000
```

割合（proportion）そのものを表示したいときは、作ったテーブルに対して `prop.table()` を使う。`prop.table()` を使うときに注意すべき点が 2 つある。まず、何を 100% にするかを決める必要がある。私たちのデータの場合、(1) 男性は男性だけで、女性は女性だけで 100% にする、(2) 内閣不支持者と内閣支持者をそれぞれ 100% にする、(3) 全体（1,000 人）を 100% にするという 3 通りが考えられる。私たちの表では、(1) の場合を**行パーセント**、(2) を**列パーセント**、(3) を**全体パーセント**という。

ここでは、「性別によって内閣支持に違いがあるか」を調べたいとする。行にある変数である性別ごとの違いを知りたいので、行パーセントを指定する（`margin = 1`）。

第 2 に、周辺度数を加える場合、行と列のうち、原因として注目していない方の周辺度数は、`prop.table()` を使う前に、注目している方の周辺度数は `prop.table()` よりあとに加えたほうがいい。私たちは性別（行変数）による違いに注目しているので、

1. 列（内閣支持）の周辺度数を加える
2. 割合を計算する
3. 行（性別）の周辺度数を加える

という順番で表を作る。

```
(tbl_fem_sup_1 <- addmargins(tbl_fem_sup, margin = 1))
```

```
        support
female   dis sup
  male   200 300
  female 250 250
  Sum    450 550
```

```
(tbl_fem_sup_p <- prop.table(tbl_fem_sup_1, margin = 1))
```

```
        support
female    dis  sup
  male   0.40 0.60
  female 0.50 0.50
  Sum    0.45 0.55
```

```
(tbl_fem_sup_2 <- addmargins(tbl_fem_sup_p, margin = 2))
```

```
        support
female    dis  sup  Sum
  male   0.40 0.60 1.00
  female 0.50 0.50 1.00
  Sum    0.45 0.55 1.00
```

```
(tbl_fem_sup_3 <- tbl_fem_sup_2 * 100)
```

```
        support
female   dis sup Sum
  male    40  60 100
  female  50  50 100
  Sum     45  55 100
```

男性の内閣支持率は 60%、女性の内閣支持率は 50% であることがわかる。

独立性の検定

私たちのデータでは、女性の内閣支持率より、男性の内閣支持率の方が高い。これはたまたま得られた結果だろうか。それとも、母集団でも男性の内閣支持率の方が高いと考えられるだろうか。

これを確かめるために、統計的検定を行う。ここで検証する仮説は以下である。

- 帰無仮説：性別と内閣支持率は独立である（つまり、男性の内閣支持率と女性の内閣支持率に差はない）
- 対立仮説：性別と内閣支持率には関連がある（つまり、男性の内閣支持率 $\neq$ 女性の内閣支持率）

この検定は、2 つの変数が独立である（帰無仮説）か、独立ではない（対立仮説）かを確かめるので、**独立性の検定**と呼ばれる。また、検定に χ^2（カイ二乗）分布を使うので、χ^2 **検定**と呼ばれることもある。

検定で使うカイ二乗分布の自由度は、分析する表の $(\text{行数} - 1)(\text{列数} - 1)$ である。周辺度数を除くと、私たちの表は 2 行$\times$2列なので、自由度$(2 - 1)(2 - 1) = 1$のカイ二乗分布を利用する。

有意水準を 7% に設定すると、検定に使う臨界値は qchisq() を使って、

```
qchisq(p = 0.07, df = 1, lower.tail = FALSE)
```

```
[1] 3.28302
```

ということがわかる。検定統計量がこの臨界値より大きいとき、私たちは帰無仮説を棄却する。反対に、検定統計量がこの臨界値以下のとき、私たちは帰無仮説を棄却しない。

検定統計量は、chisq.test() で計算できる（イェーツの連続性補正は使わないので correct = FALSE とする）。

```
chisq.test(myd$female, myd$support, correct = FALSE)
```

```

    Pearson's Chi-squared test

data:  myd$female and myd$support
X-squared = 10.101, df = 1, p-value = 0.001482
```

X-squared が検定統計量である。ここでは、10.1 という値が得られた。この値は、有意水準 7% での臨界値より大きいので、帰無仮説は棄却される。

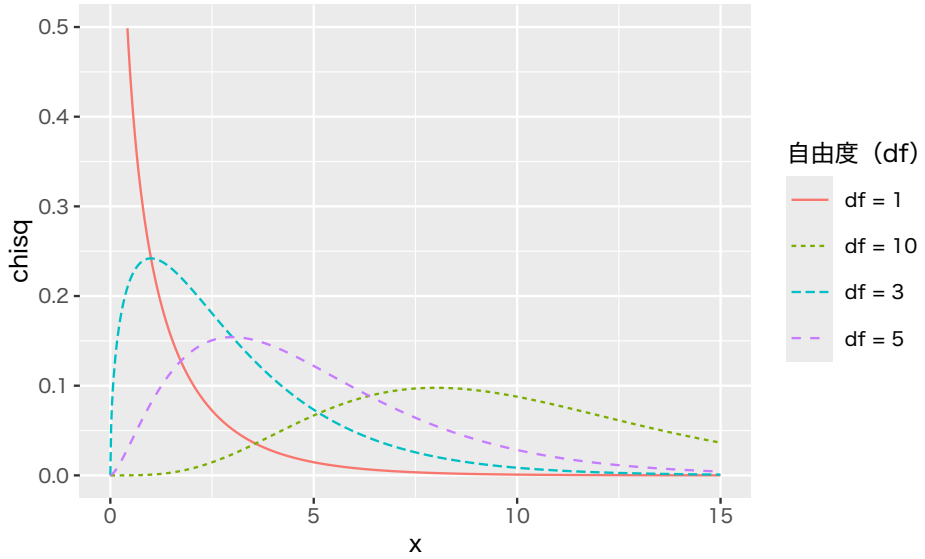
したがって、内閣支持率は性別によって異なり、男性の方が女性よりも内閣を支持するという判断を下す。

カイ二乗分布を理解する

カイ二乗（χ^2）分布は、自由度 $df \geq 1$ によってその形を変える。

```
x <- seq(0, 20, length = 1000)
chi1 <- dchisq(x, df = 1)
chi3 <- dchisq(x, df = 3)
chi5 <- dchisq(x, df = 5)
chi10 <- dchisq(x, df = 10)
df_chisq <- tibble(
  x = rep(x, 4),
  chisq = c(chi1, chi3, chi5, chi10),
  group = rep(c('df = 1', 'df = 3', 'df = 5', 'df = 10'),
              rep(1000, 4))
)
dens_chisq <- ggplot(df_chisq,
                     aes(x = x, y = chisq,
```

```
                          color    = group,
                          linetype = group)) +
  geom_line() +
  xlim(0, 15) + ylim(0, .5) +
  scale_color_discrete(name = '自由度(df)') +
  scale_linetype_discrete(name = '自由度(df)')
plot(dens_chisq)
```



量的変数同士の関係を調べる

量的変数

量的変数とは、簡単に言うと、数値自体に意味がある変数。量的変数を調べるときは、まず、基本的な統計量とヒストグラムを確認する。

```
mean(myd$height)
```

```
[1] 165.8727
```

```
median(myd$height)
```

```
[1] 166.15
```

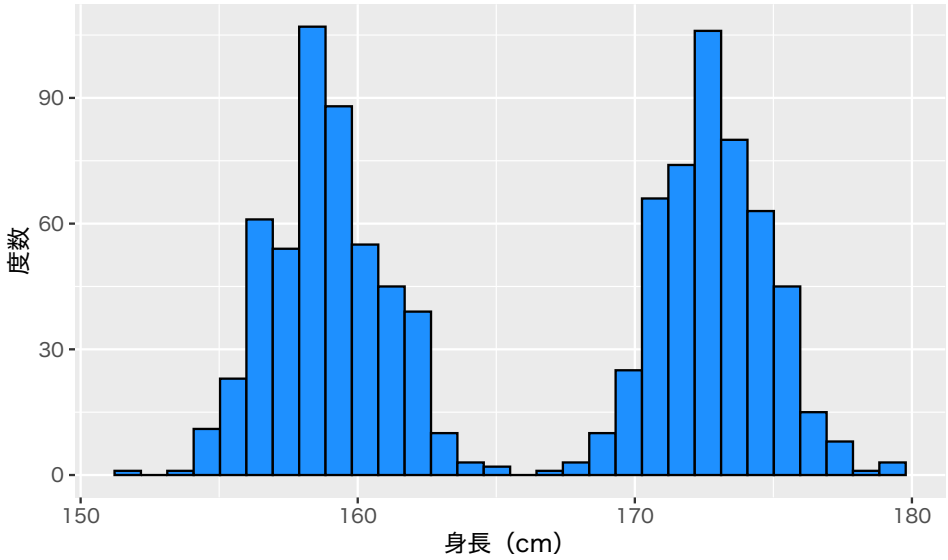
```
var(myd$height)
```

```
[1] 53.41806
```

```
sd(myd$height)
```

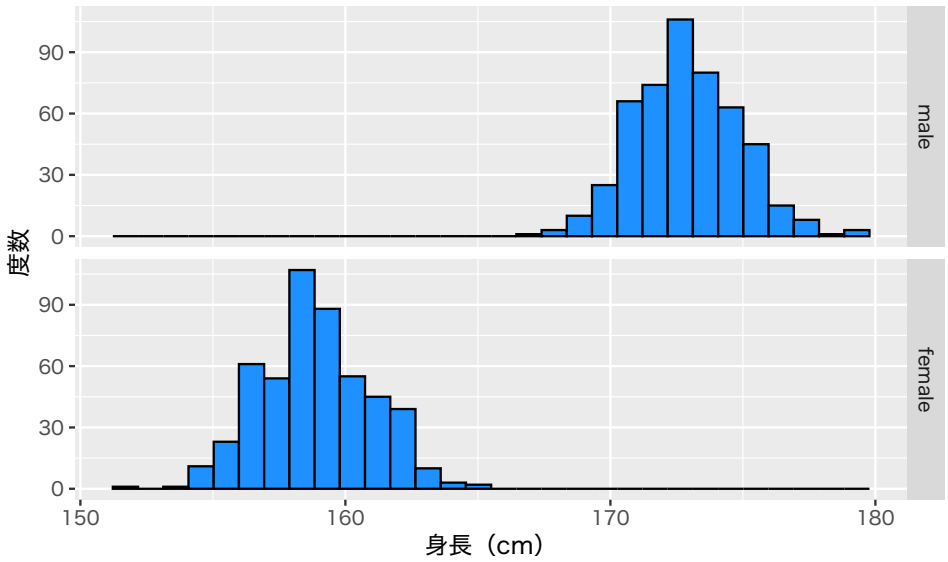
```
[1] 7.308766
```

```
h_height <- ggplot(myd, aes(x = height)) +
  geom_histogram(color = 'black', fill = 'dodgerblue') +
  labs(x = '身長(cm)', y = '度数')
plot(h_height)
```



このデータが身長のデータであり、男女ともにデータに含まれていることを考えると、性別によって分布の山が変わりそうである。男女別にしてみる。

```
h_height_gender <- h_height +
  facet_grid(row = vars(female))
plot(h_height_gender)
```



同様に、faheight の統計量を確認する。

```
mean(myd$faheight)
```

```
[1] 169.7533
```

```
median(myd$faheight)
```

```
[1] 169.6
```

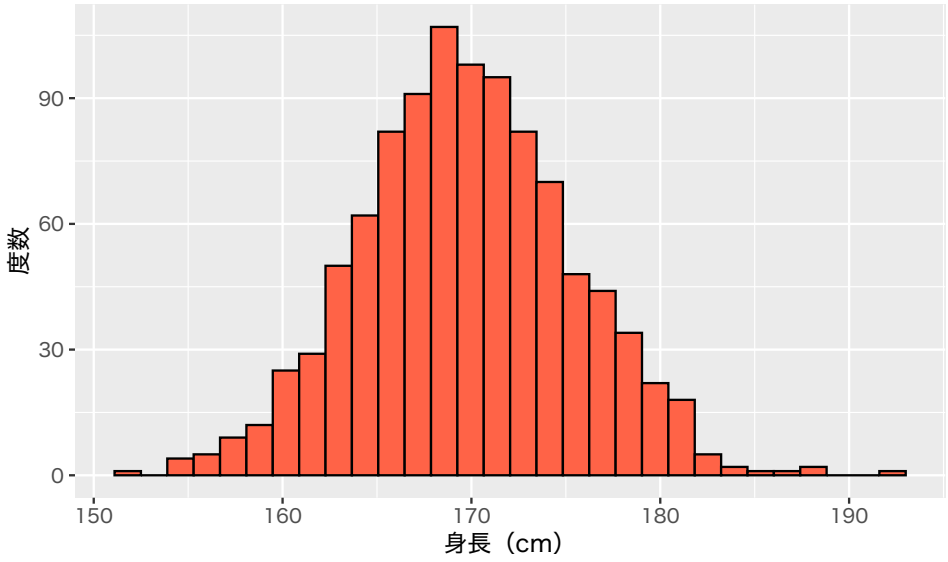
```
var(myd$faheight)
```

```
[1] 31.34405
```

```
sd(myd$faheight)
```

```
[1] 5.598576
```

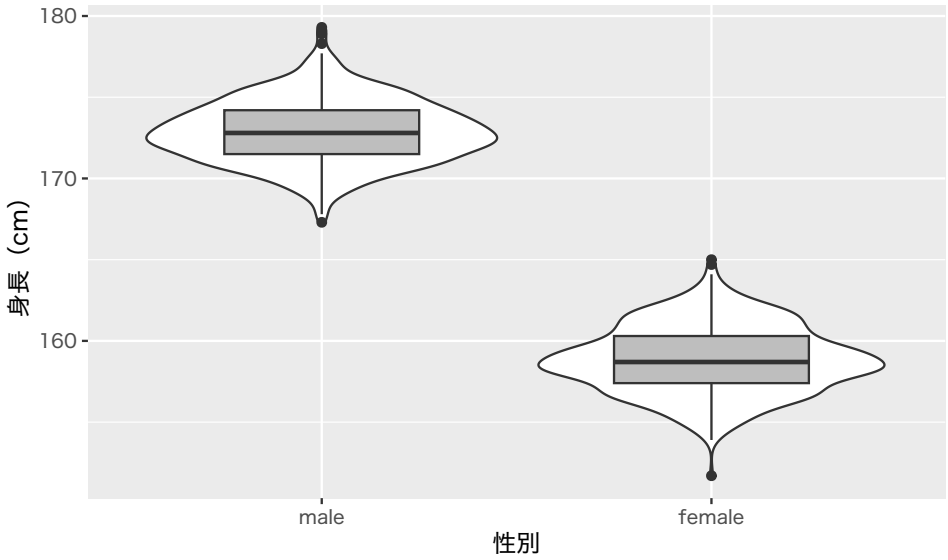
```
h_father <- ggplot(myd, aes(x = faheight)) +
  geom_histogram(color = 'black', fill = 'tomato') +
  labs(x = '身長(cm)', y = '度数')
plot(h_father)
```



質的変数と量的変数の関係

身長のヒストグラムから明らかになったように、性別（質的変数）と身長（量的変数）の間に関係がありそうである。質的変数と量的変数の関係は、図示すると理解しやすい。箱ひげ図（box-and-whisker plot, box plot）とバイオリン図（violin plot）と呼ばれる図を作って確認する。

```
vb1 <- ggplot(myd, aes(x = female, y = height)) +
  geom_violin() +
  geom_boxplot(fill = 'gray', width = 0.5) +
  labs(x = '性別', y = '身長(cm)')
plot(vb1)
```



この図を見ると、男性の身長の分布と女性の身長の分布は異なる分布であり、男性の身長のほうが高いようである。

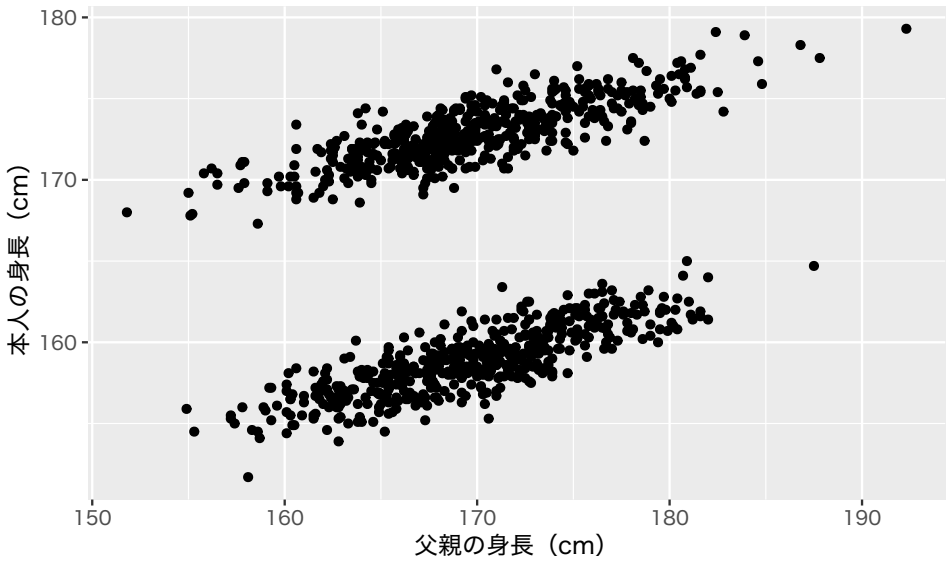
しかし、これはあくまで標本の分布であり、母集団でも同じことが言えるかどうかは検定で確かめる必要がある。これを確かめるには、平均値の差の検定（この場合はウェルチ（Welch）の t 検定）を行う必要がある。

量的変数同士の関係

次に量的変数同士（`height` と `faheight`）の関係を確かめてみる。

まず散布図を描く。

```
scat1 <- ggplot(myd, aes(x = faheight, y = height)) +
  geom_point() +
  labs(x = '父親の身長(cm)', y = '本人の身長(cm)')
plot(scat1)
```



この図を見ると、父親の身長が高いほど、子の身長が高いという関係がありそうだ。

次に、2 変数の**直線的な**関係の強さを図るために、相関係数（orrelation coefficient）を計算する。

```
with(myd, cor(height, faheight))
```

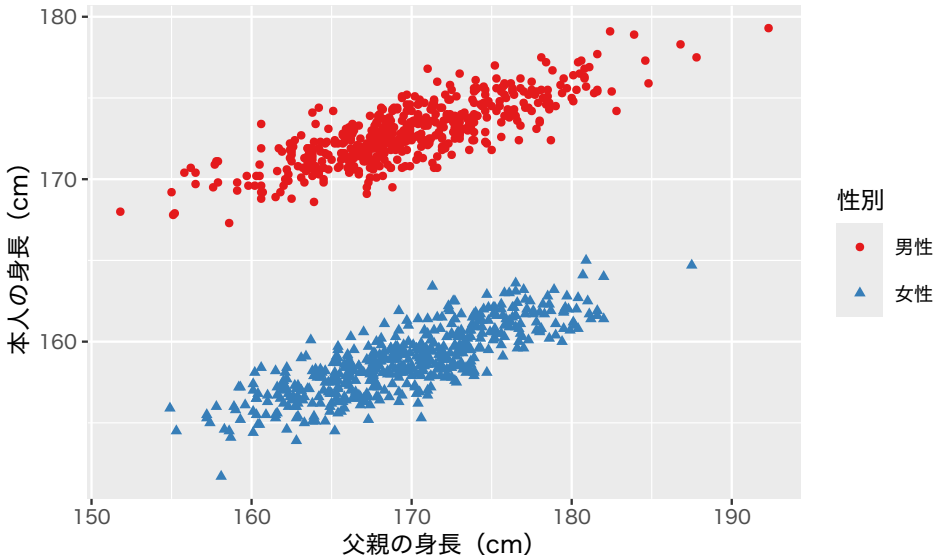
```
[1] 0.2394457
```

弱い正の相関があるようだ。

図から得られた情報は整合的だろうか。相関係数だけ見ると、父親の身長と子の身長にはあまり強い関係は無さそうだという結論に成る。しかし、散布図を見ると、2 変数の間には非常に強い関係がありそうである。

男女を色と形で区別して散布図を作り直してみる。

```
scat2 <- ggplot(myd, aes(x = faheight, y = height,
                         color = female,
                         shape = female)) +
  geom_point() +
  labs(x = '父親の身長(cm)', y = '本人の身長(cm)') +
  scale_color_brewer(palette = 'Set1',
                     name   = '性別',
                     labels = c('男性', '女性')) +
  scale_shape_discrete(name   = '性別',
                       labels = c('男性', '女性'))
plot(scat2)
```



性別ごとにグループができている。

男女別にして相関係数を求め直そう。

```
myd |>
  filter(female == 'female') |>
  with(cor(height, faheight)) |>
  round(digits = 2)
```

```
[1] 0.81
```

```
myd |>
  filter(female == 'male') |>
  with(cor(height, faheight)) |>
  round(2)
```

```
[1] 0.81
```

この例から分かる通り、単に相関係数だけを求めると、2 変数の関係を見誤る可能性がある。反対に、散布図だけに頼ると、ありもしないパタンを「見つけて」しまうことがある。