Rで実践！美しいTable & Figureを作ろう

Part 3: 推測統計の可視化

例：生存時間解析

佐藤 俊太朗
長崎大学病院 臨床研究センター
shuntarosato@nagasaki-u.ac.jp
Twitter: @Shuntarooo3

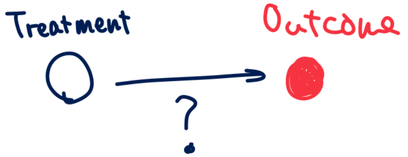
本パートのテーマ

因果推論

処置がアウトカムに
影響するか？

喫煙の程度は、死亡に影響するか？

NHEFSデータ（観察研究データ）で評価する



曝露変数

喫煙の程度

▼

Pack-years

アウトカム変数

死亡

▼

(死亡イベント or 打ち切り) + 観察時間

▼

生存時間データ　より一般的には、time-to-eventデータ

▼ 生存時間データを解析するので

生存時間解析

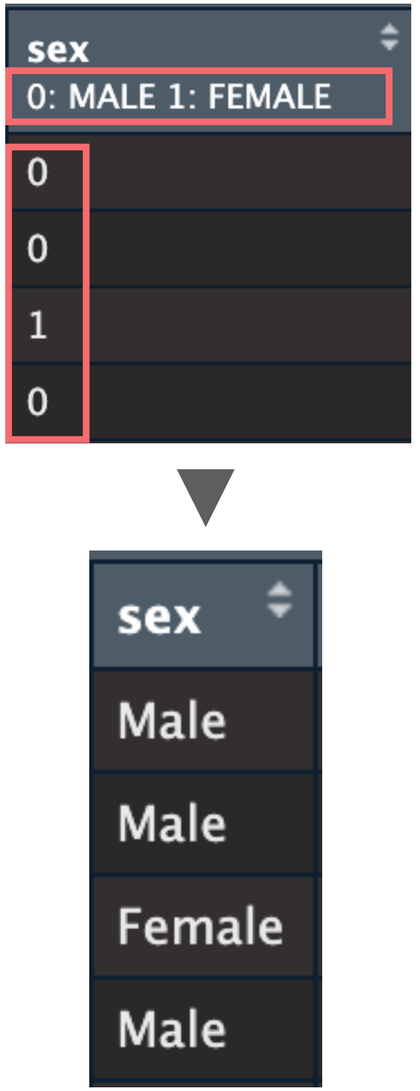
データセットを作ろう

Package（HP: 1）

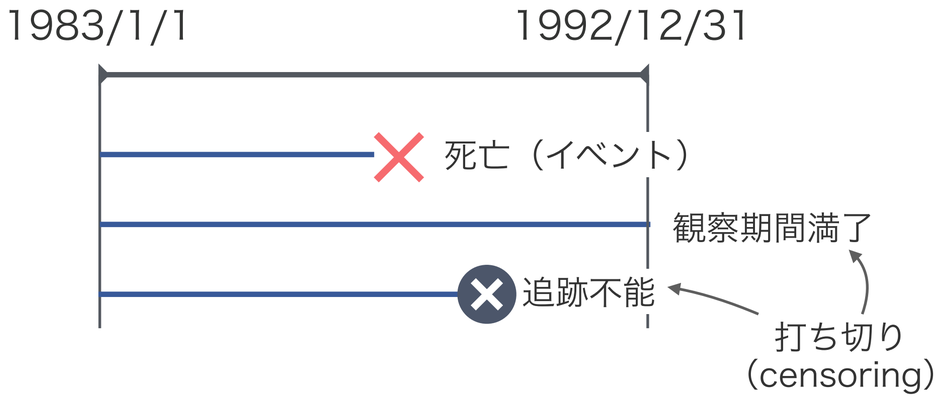
```
# install.packages("pacman")
library(pacman)
p_load(tidyverse,  # データセットのハンドリン
グ      lubridate,  # 時間データのハンドリング
        causaldata,  # データセット集
        labelled,  # 変数のラベルの調整
        ggdag,  # DAGを描く
        gtsummary,  # データの要約
        survival,  # 生存時間解析
        ggsurvfit,  # 生存時間解析
        WeightIt,  # 逆確率重み付け
        cobalt,  # 標準化差
        broom  # 結果の整形
        )
```

データセット（HP: 2）

```r
data(nhefs_complete)
var_label(nhefs_complete) <- NULL  # nhefs_complete に入っている変数説明情報を削除
df01 <- nhefs_complete |>
     # 変数選択
     select(seqn, death, yrdth, modth, dadth, sex, age, race, education,
            exercise, smokeintensity, smokeyrs) |>
     # カテゴリー化、ラベリング
     mutate(
          sex = factor(sex,
                       labels = c("Male", "Female")),
          race = factor(race,
                        labels = c("White", "Black or other")),
          education = factor(education,
                             labels = c("8th grade or less",
                                        "HS dropout",
                                        "HS",
                                        "College dropout",
                                        "College or more")),
          exercise = factor(exercise,
                            labels = c("Much exercise",
                                       "Moderate exercise",
                                       "Little or no exercise"))
     )
```



アウトカム変数（生存時間データ）（HP: 4.1）



生存時間データ

- 死亡の有無
- 観察できた時間

```
df02 <- df01 |>
     mutate(
          # 死亡日または打ち切り日作成
          event_date = if_else(death == 0, ymd("1992-12-31"),
                               ymd(19000000 + yrdth * 10000 + modth * 100 + dadth)
                               ),
          # 観察開始日からの日数を計算
          survtime = difftime(time1 = event_date,
                              time2 = ymd("1983-01-01")
                              )
          )
```

death	yrdth	modth	dadth
0	*NA*	*NA*	*NA*
1	85	2	14

event_date	survtime
1992-12-31	3652 days
1985-02-14	775 days

曝露変数（Pack-years）（HP: 4.2）

$$\text{Pack-years} = \frac{\text{1日の喫煙本数}}{\text{20本}} \times \text{喫煙年数}$$

```
df03 <- df02 |>
     mutate(
          pack_years_n = (smokeintensity / 20) * smokeyrs
     )

df04 <- df03 |>
     mutate(
          pack_years = if_else(pack_years_n >= 20, 1, 0),
          pack_years = factor(pack_years,
                              labels = c("Low", "High"))
     )
```

smokeintensity	smokeyrs
30	29
20	24

pack_years_n	pack_years
24.00	High
26.00	High
7.95	Low
19.00	Low

おまじない（打ち切りの作成と死亡イベントの水増し）（HP: 4.3）

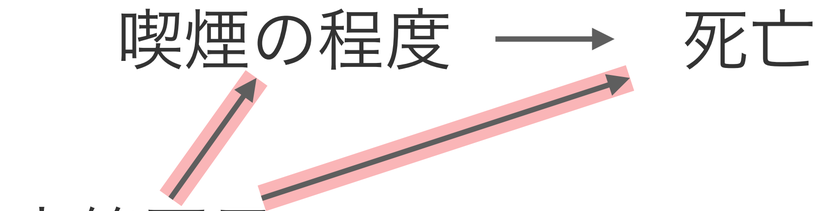
このコードの理解は不要です

```
set.seed(1234)

df05 <- df04 |>
     mutate(death1 = rbinom(1566, 1, 0.6),
            death0 = rbinom(1566, 1, 0.4),
            death2 = if_else(pack_years == "High", death1, death0),
            censor = rbinom(1566, 1, 0.2),
            r_time = if_else(censor == 1 & death == 0 | death2 == 1 & death == 0,
                             round(runif(1566, min = 0, 3652)), 0),
            death = if_else(death2 == 1, 1, death),
            survtime = survtime - r_time,
            survtime_y = as.numeric(survtime) / 365.25)
```

因果構造の確認

交絡と交絡因子

喫煙の程度は、死亡に影響するか？



- sex
- age
- race
- education
- exercise

交絡因子があると...

データから直接計算できる

喫煙の程度と死亡との関連　≠　喫煙の程度の死亡への影響

違う→
交絡因子によるバイアス（交絡）

▼

交絡因子を統制すると

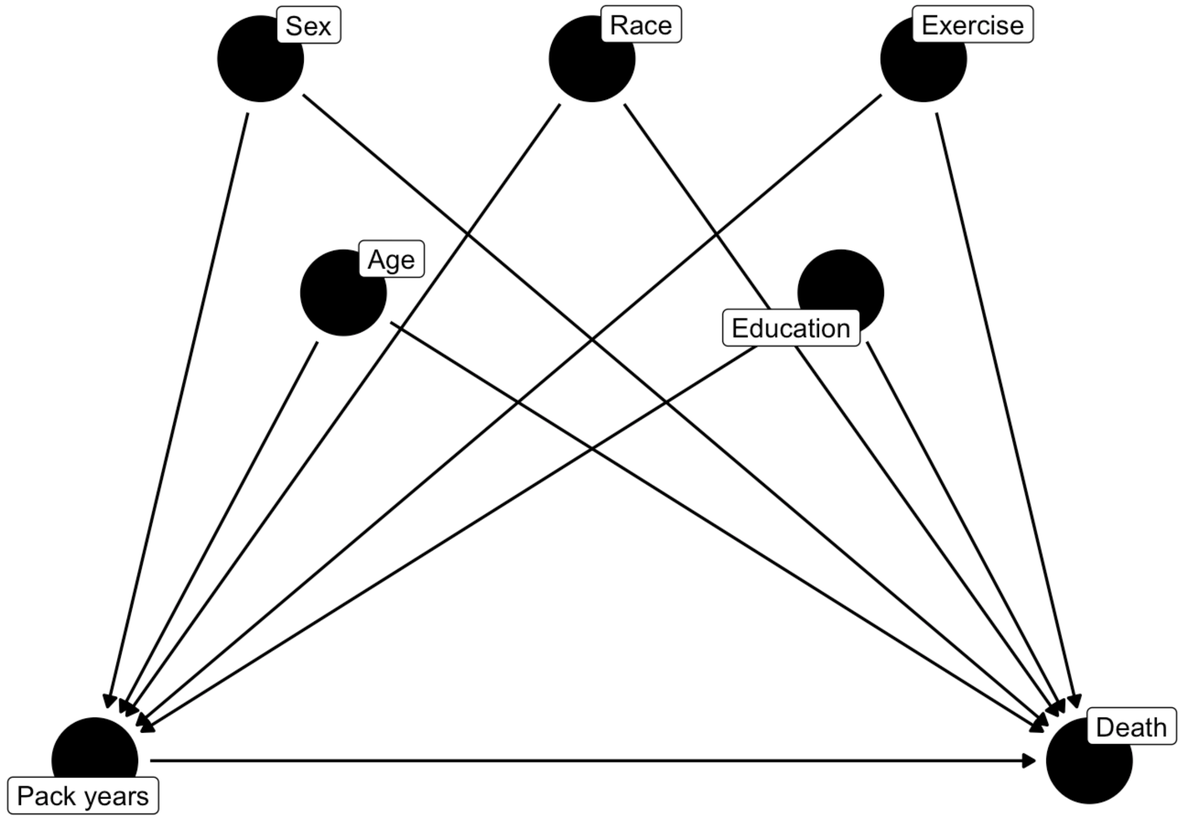
データから直接計算できる

喫煙の程度と死亡との関連　＝　喫煙の程度の死亡への影響

等しい

DAGで因果構造を図示する（HP: 4.4）

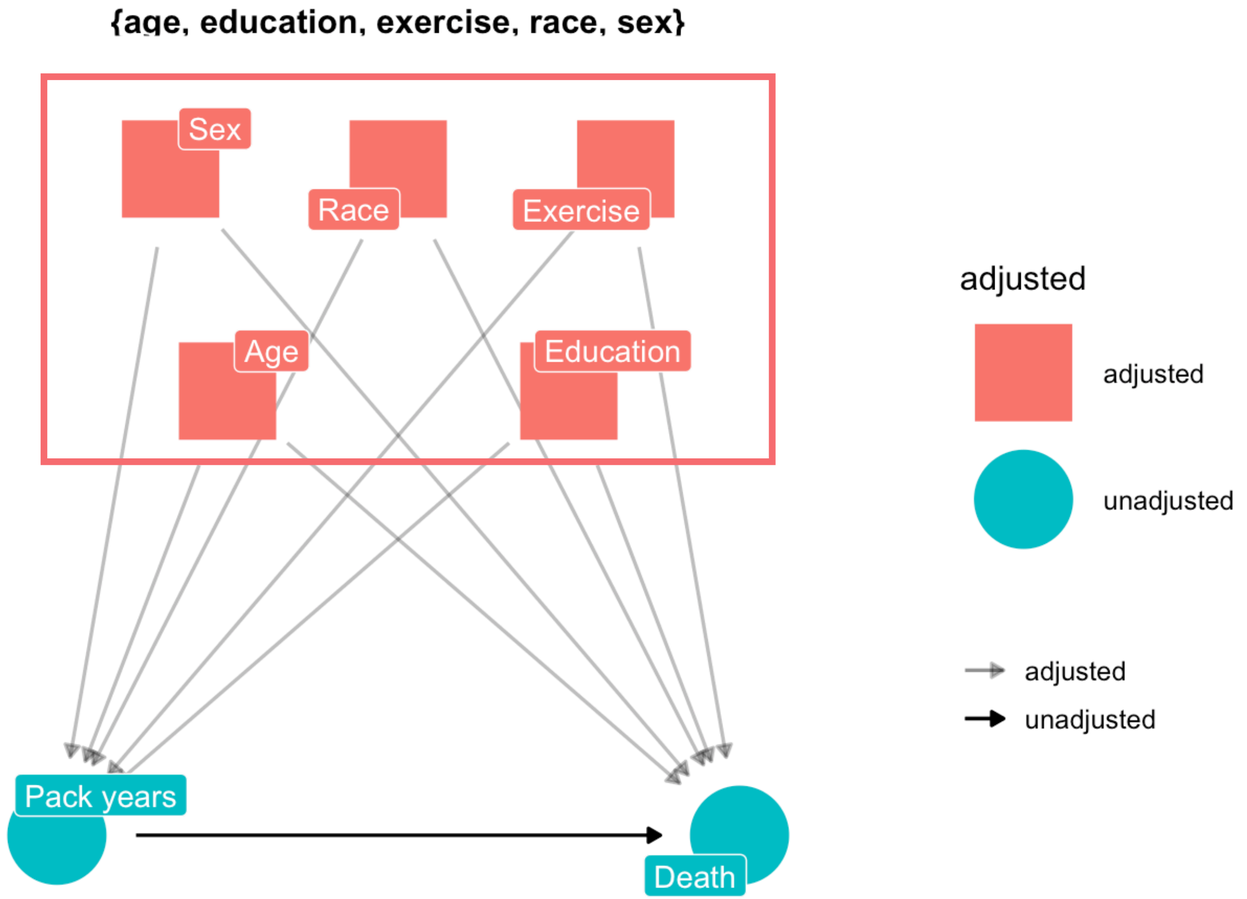
```
dag01 <- dagify(
     # 因果構造の指定
     death ~ pack_years + sex + age + race + education + exercise,
     pack_years ~ sex + age + race + education + exercise,
     exposure = "pack_years",
     outcome = "death",
     # 変数の位置指定
     coords = list(
          x = c(pack_years = 0,
                sex = 1,
                age = 1.5,
                race = 3,
                education = 4.5,
                exercise = 5,
                death = 6),
          y = c(pack_years = 0,
                sex = 1.5,
                age = 1,
                race = 1.5,
                education = 1,
                exercise = 1.5,
                death = 0)
          ),
     # 変数のラベリング
     labels = c(
               "pack_years" = "Pack years",
               "sex" = "Sex",
               "age" = "Age",
               "race" = "Race",
               "education" = "Education",
               "exercise" = "Exercise",
               "death" = "Death")
     ) |>
     tidy_dagitty()

ggdag(dag01,
      text = FALSE,
      use_labels = "label") +
     theme_dag()
```



Rで交絡因子を見つける（HP: 4.4）

ggdagパッケージでは、
バックドア基準を使って、
交絡因子を見つけることができる

```
ggdag_adjustment_set(dag01,
       text = FALSE,
       use_labels = "label",
       shadow = TRUE) +
      theme_dag()
```



背景情報の要約（Table 1）

Table 1

- 研究者は自分がもともと想定していた集団か考えられる
- 読者は目の前の患者さんや集団に研究の結果を適用できるか考えられる

地味で、作るのは面倒

▼

Rなら簡単にできます！！

Table 1. Participant Demographics, Medical History, and Medications at Baseli With Major Ischemic or Hemorrhagic Events in Atrial Fibrillation[a,b]

	Unweighted	
Characteristic	Rivaroxaban, %	Apixaban, %
No. of participants	227 572	353 879
Anticoagulant dose reduced	23.0	23.2
Demographics		
Age, mean (SD), y	76.3 (6.8)	77.4 (7.2)
Year anticoagulant started, mean	2015.5	2016.3
Women	48.3	51.4
Men	51.7	48.6
Race and ethnicity[c]		
Asian	1.6	1.3
Black	3.6	3.7
Hispanic	1.2	1.0
North American Native	0.3	0.2
White	92.0	92.6
Other	1.2	1.1

Ray WA, Chung CP, Stein CM, Smalley W, Zimmerman E, Dupont WD, et al. Association of Rivaroxaban vs Apixaban With Major Ischemic or Hemorrhagic Events in Patients With Atrial Fibrillation. JAMA. 2021 Dec 21;326(23):2395–404.

gtsummaryパッケージで表を作る（デフォルト）（HP: 5.1）

```
df05 |>
     select(pack_years,  # 曝露変数
            sex, age, race, education, exercise  # 背景情報
            ) |>
     tbl_summary( )
```

- Pack-yearsの群（High, Low）ごとに要約して欲しい
- （他に良い方法があるが）群間比較をして欲しい

Characteristic	**N = 1,566**[1]
pack_years	
Low	744 (48%)
High	822 (52%)
sex	
Male	762 (49%)
Female	804 (51%)
age	43 (33, 53)
race	
White	1,360 (87%)
Black or other	206 (13%)
education	
8th grade or less	291 (19%)
HS dropout	340 (22%)
HS	637 (41%)
College dropout	121 (7.7%)
College or more	177 (11%)
exercise	
Much exercise	300 (19%)
Moderate exercise	661 (42%)
Little or no exercise	605 (39%)

[1] n (%); Median (IQR)

完成版Table 1を作ろう！（HP: 5.1）

前のページのコードを工夫して
右の表を作りましょう

- 群ごとの要約
  - tbl_summary関数のby引数を使う
- 群間比較
  - add_p()を使う

Characteristic	Low, N = 744[1]	High, N = 822[1]	p-value[2]
sex			<0.001
Male	282 (38%)	480 (58%)	
Female	462 (62%)	342 (42%)	
age	34 (29, 46)	48 (42, 56)	<0.001
race			<0.001
White	602 (81%)	758 (92%)	
Black or other	142 (19%)	64 (7.8%)	
education			<0.001
8th grade or less	105 (14%)	186 (23%)	
HS dropout	133 (18%)	207 (25%)	
HS	331 (44%)	306 (37%)	
College dropout	70 (9.4%)	51 (6.2%)	
College or more	105 (14%)	72 (8.8%)	
exercise			0.075
Much exercise	160 (22%)	140 (17%)	
Moderate exercise	308 (41%)	353 (43%)	
Little or no exercise	276 (37%)	329 (40%)	

[1] n (%); Median (IQR)
[2] Pearson's Chi-squared test; Wilcoxon rank sum test

完成版Table 1を作ろう！（回答）（HP: 5.1）

```
df05 |>
     select(pack_years,  # 曝露変数
            sex, age, race, education, exercise  # 背景情報
            ) |>
     tbl_summary(by = pack_years) |>  # 群間比較にする
     add_p()
```

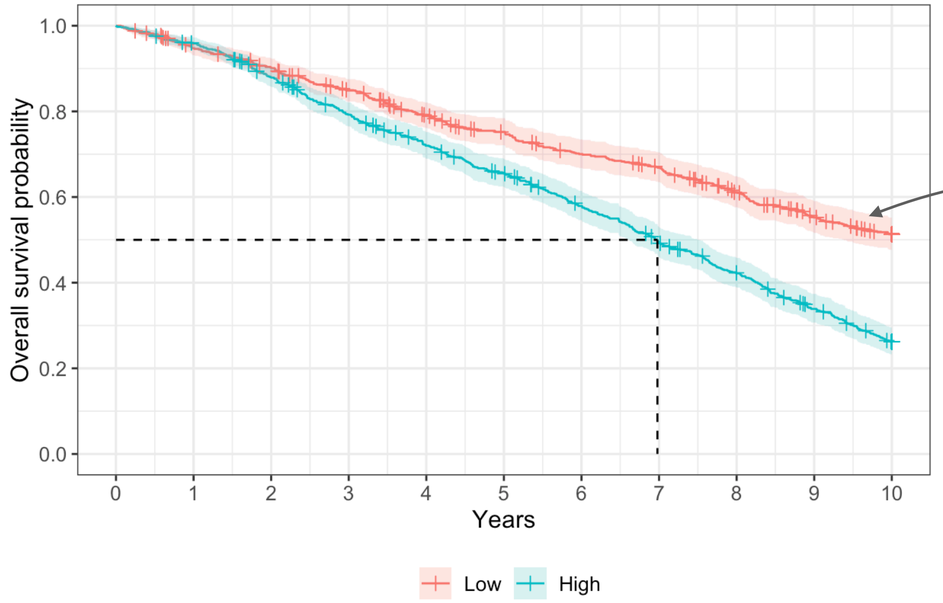
Characteristic	Low, N = 744[1]	High, N = 822[1]	p-value[2]
sex			<0.001
Male	282 (38%)	480 (58%)	
Female	462 (62%)	342 (42%)	
age	34 (29, 46)	48 (42, 56)	<0.001
race			<0.001
White	602 (81%)	758 (92%)	
Black or other	142 (19%)	64 (7.8%)	
education			<0.001
8th grade or less	105 (14%)	186 (23%)	
HS dropout	133 (18%)	207 (25%)	
HS	331 (44%)	306 (37%)	
College dropout	70 (9.4%)	51 (6.2%)	
College or more	105 (14%)	72 (8.8%)	
exercise			0.075
Much exercise	160 (22%)	140 (17%)	
Moderate exercise	308 (41%)	353 (43%)	
Little or no exercise	276 (37%)	329 (40%)	

[1] n (%); Median (IQR)
[2] Pearson's Chi-squared test; Wilcoxon rank sum test

アウトカムの分布の確認

Kaplan-Meier曲線（HP: 5.2）

生存時間データの分布の確認には、Kaplan-Meier推定に基づいた生存確率を図示することが多い



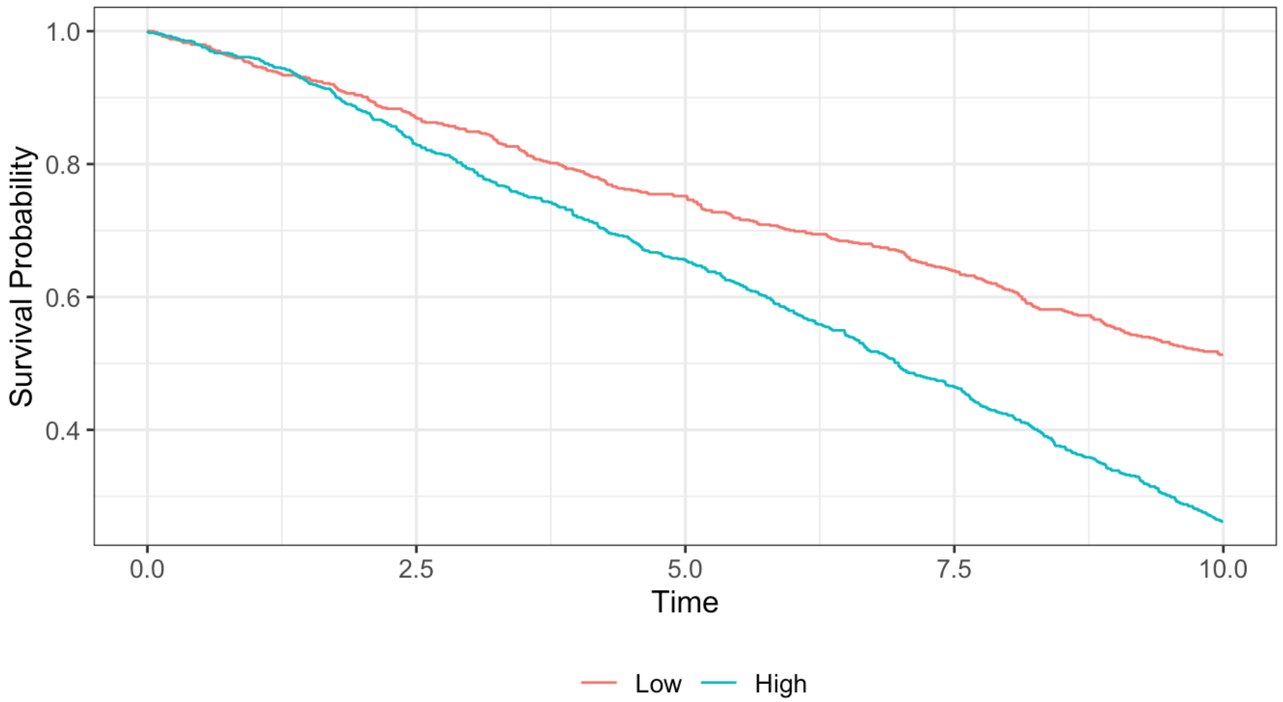
打ち切り
（ひげと言う）

No. at risk表

Low											
At Risk	744	697	661	613	561	525	486	460	409	358	0
Events	0	39	72	111	153	180	216	238	277	316	341
Censored	0	8	11	20	30	39	42	46	58	70	403
High											
At Risk	822	785	715	638	574	519	452	384	323	254	0
Events	0	34	98	169	227	278	338	404	460	523	580
Censored	0	3	9	15	21	25	32	34	40	45	242

Kaplan-Meier曲線を描こう（デフォルト設定）（HP: 5.2）

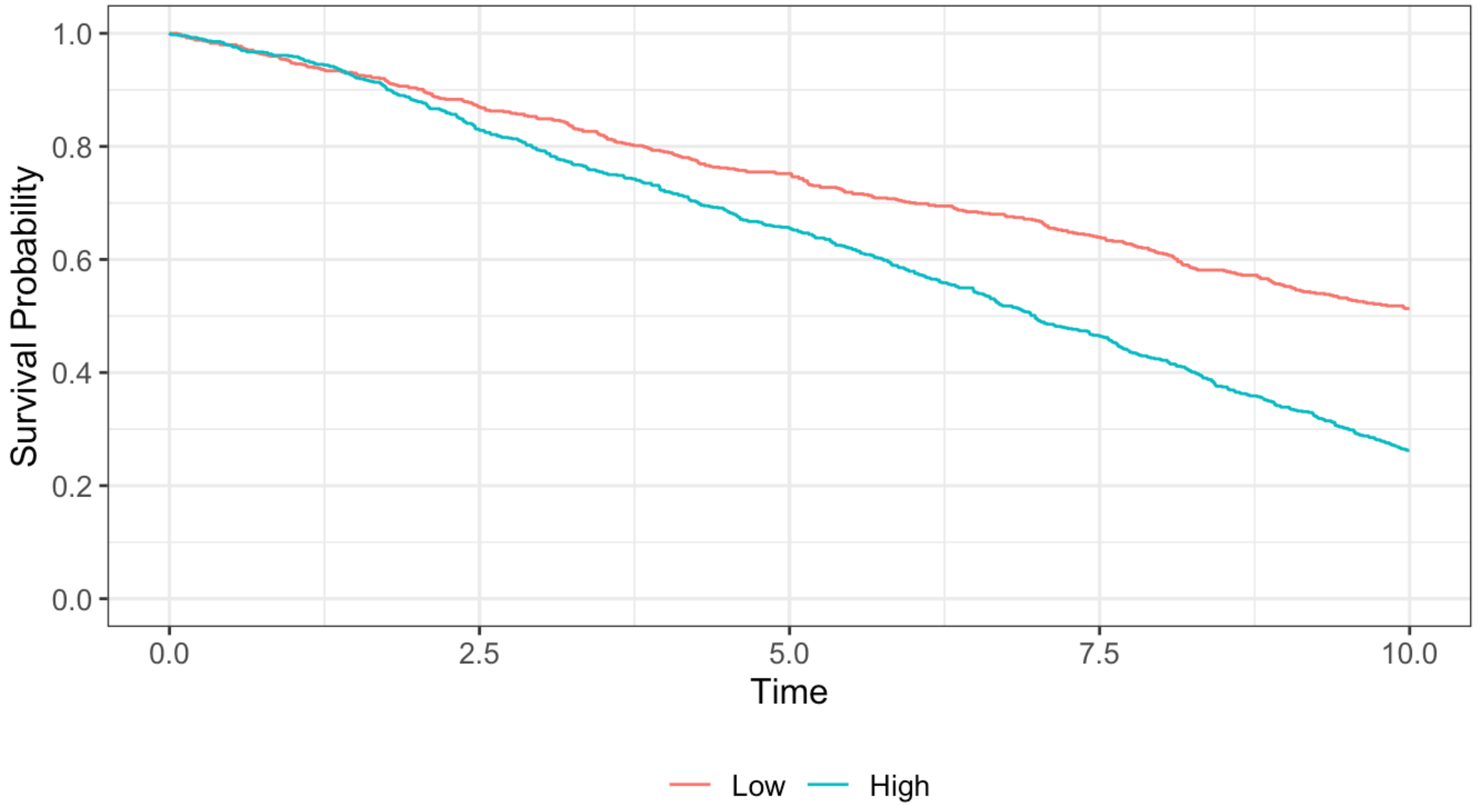
```r
# Kaplan-Meier推定
# Kaplan-Meier方に基づく生存確率を推定
km_simple <- survfit2(Surv(survtime_y, death) ~ pack_years, data = df05)

ggsurvfit(km_simple)
```



- Y軸が確率なのに、
  0から始まっていない
- 軸タイトルをわかりやすくしたい
- 95%信頼区間が欲しい
- 打ち切りを示したい
  - ひげ
  - No. at risk表
- 生存時間中央値を示したい

Kaplan-Meier曲線を描こう（軸の設定）（HP: 5.2）

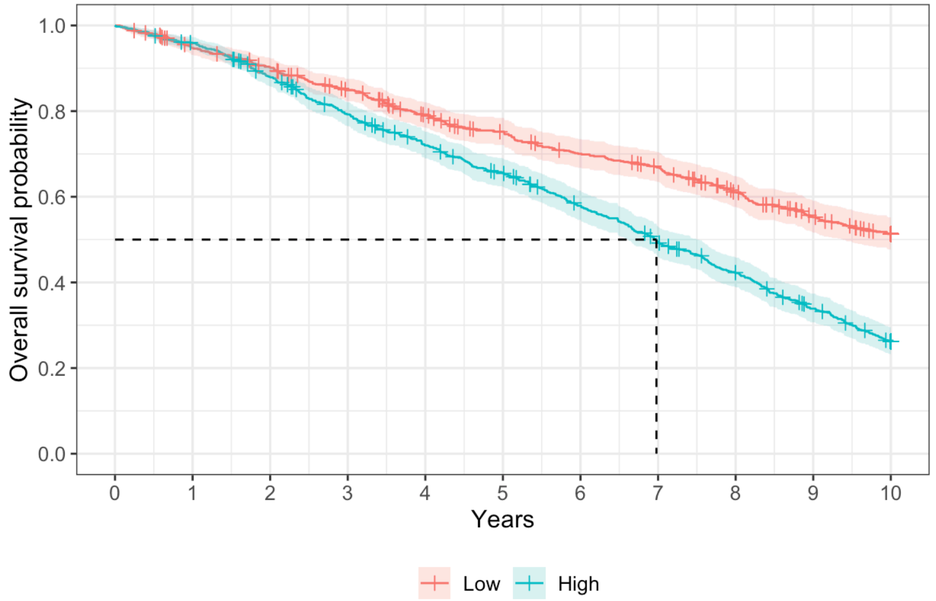
```
ggsurvfit(km_simple) +
     scale_y_continuous(limits = c(0, 1),
                        breaks = seq(0, 1, by = 0.2))
```



完成版Kaplan-Meier曲線を描こう（HP: 5.2）

前のページのコードを工夫して
右の図を作りましょう

- X軸とY軸に任意のタイトルをいれる
  （labs(x = "Years", y = "Overall survival probability")を追加する）
- Kaplan-Meier曲線に95%信頼区間をつける
  （add_confidence_interval()を追加する）
- 打ち切り（ひげ）をつける
  （add_censor_mark()を追加する）
- No. at risk表をつける
  （add_risktable(risktable_stats = c("n.risk", "cum.event", "cum.censor"))を追加する）
- 生存時間中央値を示す
  （add_quantile()を追加する）
- X軸の目盛りをわかりやすくする
  （scale_x_continuous(limits = c(0, 10), breaks = seq(0, 10, by = 1))を追加する）



Low											
At Risk	744	697	661	613	561	525	486	460	409	358	0
Events	0	39	72	111	153	180	216	238	277	316	341
Censored	0	8	11	20	30	39	42	46	58	70	403
High											
At Risk	822	785	715	638	574	519	452	384	323	254	0
Events	0	34	98	169	227	278	338	404	460	523	580
Censored	0	3	9	15	21	25	32	34	40	45	242

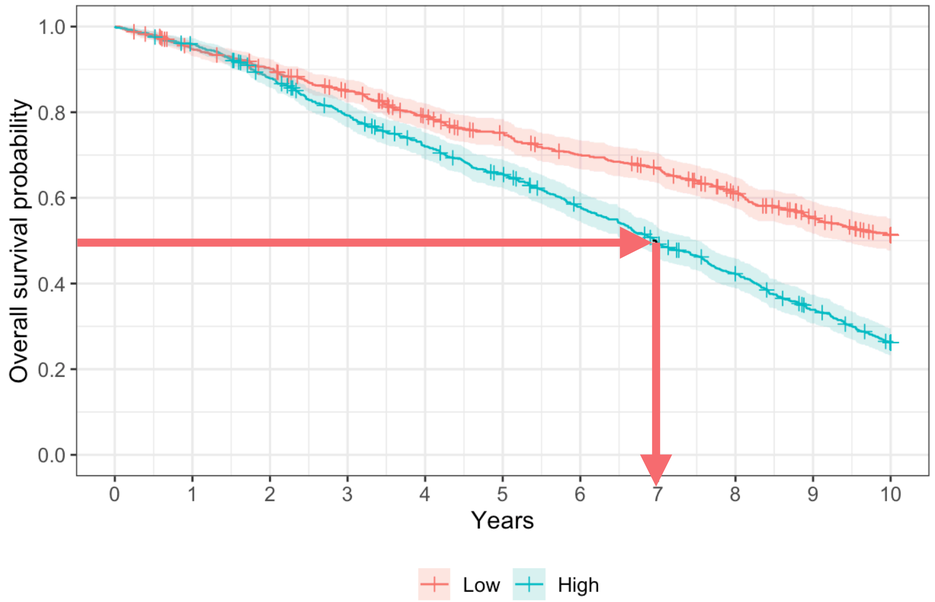
喫煙の程度と死亡との**関連**評価

生存時間中央値（MST: Median Survival Time）（HP: 5.3.1）

生存確率が50%である生存時間の推定値

```
km_simple
```

```
Call: survfit(formula = Surv(survtime_y, death) ~ pack_years, data = df05)

                  n events median 0.95LCL 0.95UCL
pack_years=Low  744    341     NA    9.38      NA
pack_years=High 822    580   6.98    6.62    7.51
```



Low											
At Risk	744	697	661	613	561	525	486	460	409	358	0
Events	0	39	72	111	153	180	216	238	277	316	341
Censored	0	8	11	20	30	39	42	46	58	70	403

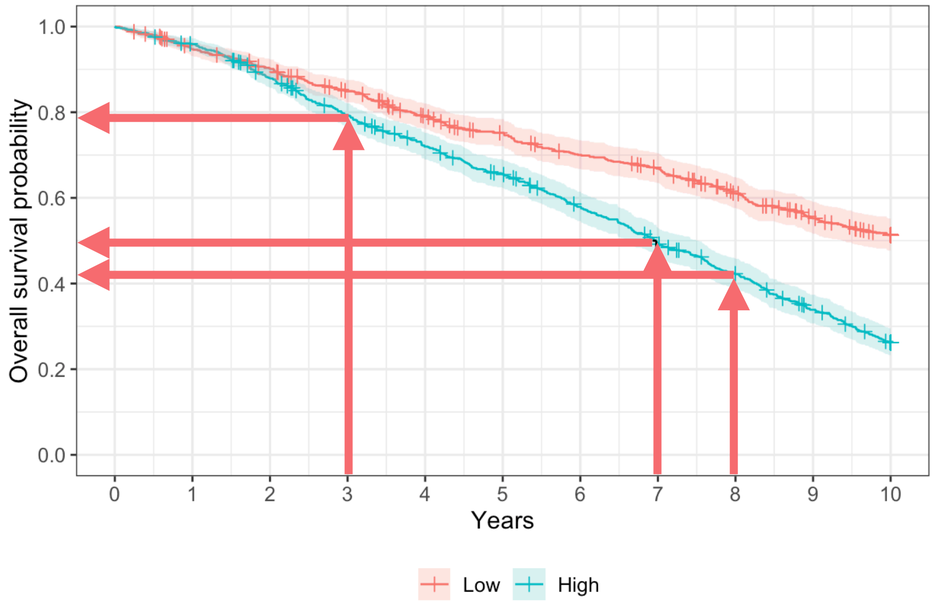
High											
At Risk	822	785	715	638	574	519	452	384	323	254	0
Events	0	34	98	169	227	278	338	404	460	523	580
Censored	0	3	9	15	21	25	32	34	40	45	242

群の時点ごとの生存確率と95%信頼区間（HP: 5.3.2）

時点ごとの生存確率の推定値

```
summary(km_simple, times = seq(0,10, by =1))
```

```
Call: survfit(formula = Surv(survtime_y, death) ~ pack_years, data = df05)

                pack_years=Low
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    0    744       0    1.000 0.00000        1.000        1.000
    1    697      39    0.947 0.00822        0.931        0.964
    2    661      33    0.902 0.01093        0.881        0.924
    3    613      39    0.849 0.01323        0.823        0.875
    4    561      42    0.790 0.01509        0.761        0.820
    5    525      27    0.752 0.01606        0.721        0.784
    6    486      36    0.700 0.01711        0.667        0.735
    7    460      22    0.668 0.01762        0.635        0.704
    8    409      39    0.611 0.01834        0.576        0.648
    9    358      39    0.552 0.01885        0.516        0.590

                pack_years=High
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    0    822       0    1.000 0.00000        1.000        1.000
    1    785      34    0.959 0.00695        0.945        0.972
    2    715      64    0.880 0.01136        0.858        0.903
    3    638      71    0.792 0.01423        0.765        0.821
    4    574      58    0.720 0.01579        0.690        0.751
    5    519      51    0.656 0.01674        0.624        0.689
    6    452      60    0.579 0.01746        0.546        0.614
    7    384      66    0.495 0.01775        0.461        0.531
    8    323      56    0.422 0.01761        0.389        0.458
    9    254      63    0.339 0.01696        0.307        0.374
```



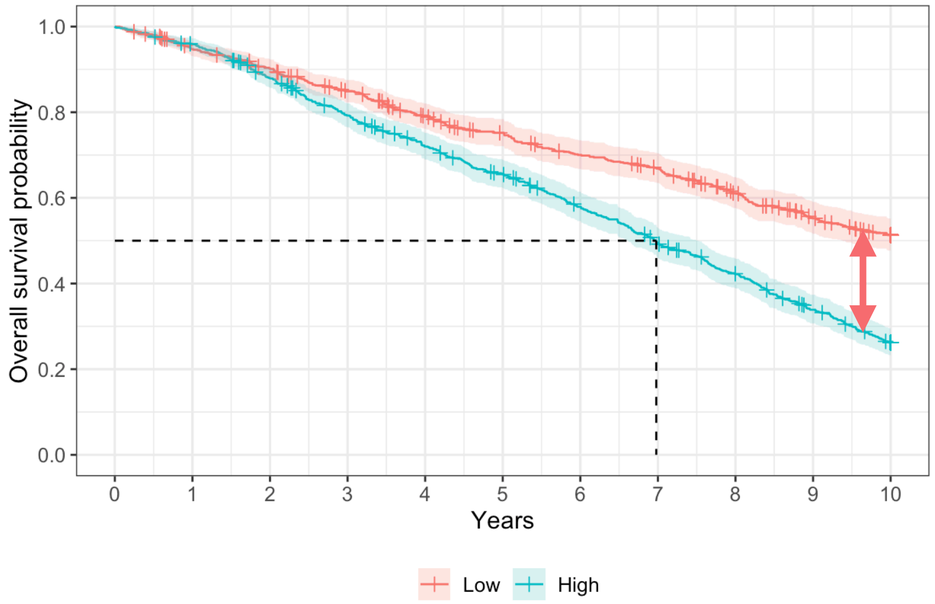
Low											
At Risk	744	697	661	613	561	525	486	460	409	358	0
Events	0	39	72	111	153	180	216	238	277	316	341
Censored	0	8	11	20	30	39	42	46	58	70	403

High											
At Risk	822	785	715	638	574	519	452	384	323	254	0
Events	0	34	98	169	227	278	338	404	460	523	580
Censored	0	3	9	15	21	25	32	34	40	45	242

生存確率の群間比較（Logrank検定）（HP: 5.3.3）

```
survdiff(Surv(survtime_y, death) ~ pack_years, data = df05)
```

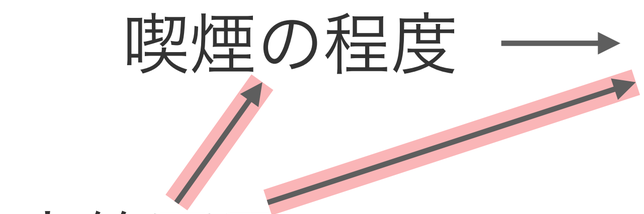
```
Call:
survdiff(formula = Surv(survtime_y, death) ~ pack_years, data = df05)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
pack_years=Low  744      341      477      38.9      81.4
pack_years=High 822      580      444      41.9      81.4

 Chisq= 81.4  on 1 degrees of freedom, p= <2e-16
```



Low											
At Risk	744	697	661	613	561	525	486	460	409	358	0
Events	0	39	72	111	153	180	216	238	277	316	341
Censored	0	8	11	20	30	39	42	46	58	70	403
High											
At Risk	822	785	715	638	574	519	452	384	323	254	0
Events	0	34	98	169	227	278	338	404	460	523	580
Censored	0	3	9	15	21	25	32	34	40	45	242

喫煙の程度の死亡への**影響**評価

交絡の調整方法



- sex
- age
- race
- education
- exercise

▶

- 多変量回帰モデル
- 傾向スコア
  - マッチング
  - 逆確率重み付け
    （IPW, IPTW: Inverse probability treatment weight）

本テーマでは、集団全体が喫煙量を減らすと予後はどうなるか、といったことに関心がある

▼

標的集団は集団全体であり、知りたい効果は平均処置効果

ATE: Average treatment effect

▼

だから、逆確率重み付けを使った交絡調整を選択

群間のバランス

Characteristic	Low, N = 744[1]	High, N = 822[1]	p-value[2]
sex			<0.001
Male	282 (38%)	480 (58%)	
Female	462 (62%)	342 (42%)	
age	34 (29, 46)	48 (42, 56)	<0.001
race			<0.001
White	602 (81%)	758 (92%)	
Black or other	142 (19%)	64 (7.8%)	
education			<0.001
8th grade or less	105 (14%)	186 (23%)	
HS dropout	133 (18%)	207 (25%)	
HS	331 (44%)	306 (37%)	
College dropout	70 (9.4%)	51 (6.2%)	
College or more	105 (14%)	72 (8.8%)	
exercise			0.075
Much exercise	160 (22%)	140 (17%)	
Moderate exercise	308 (41%)	353 (43%)	
Little or no exercise	276 (37%)	329 (40%)	

[1] n (%); Median (IQR)
[2] Pearson's Chi-squared test; Wilcoxon rank sum test

ずれてる

▼

交絡になる

▼

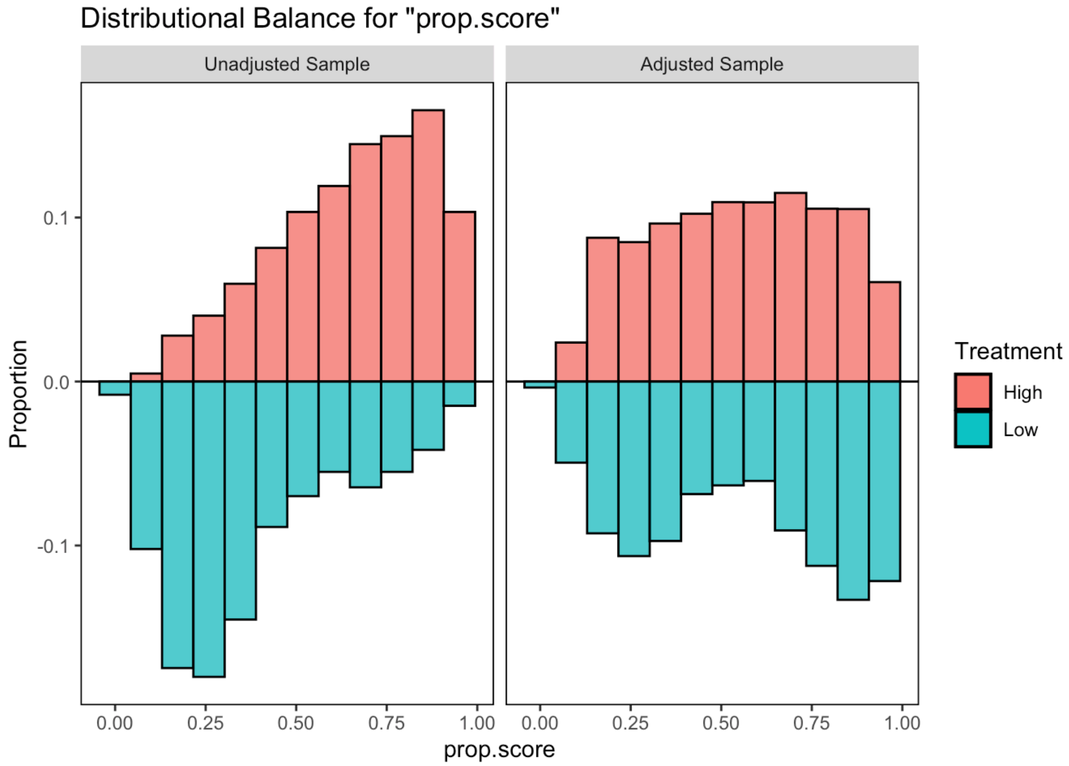
IPWで調整する

▼

傾向スコアの分布や
バランスを確認

WeightItパッケージを使った傾向スコア分布の確認（HP: 5.4.2）

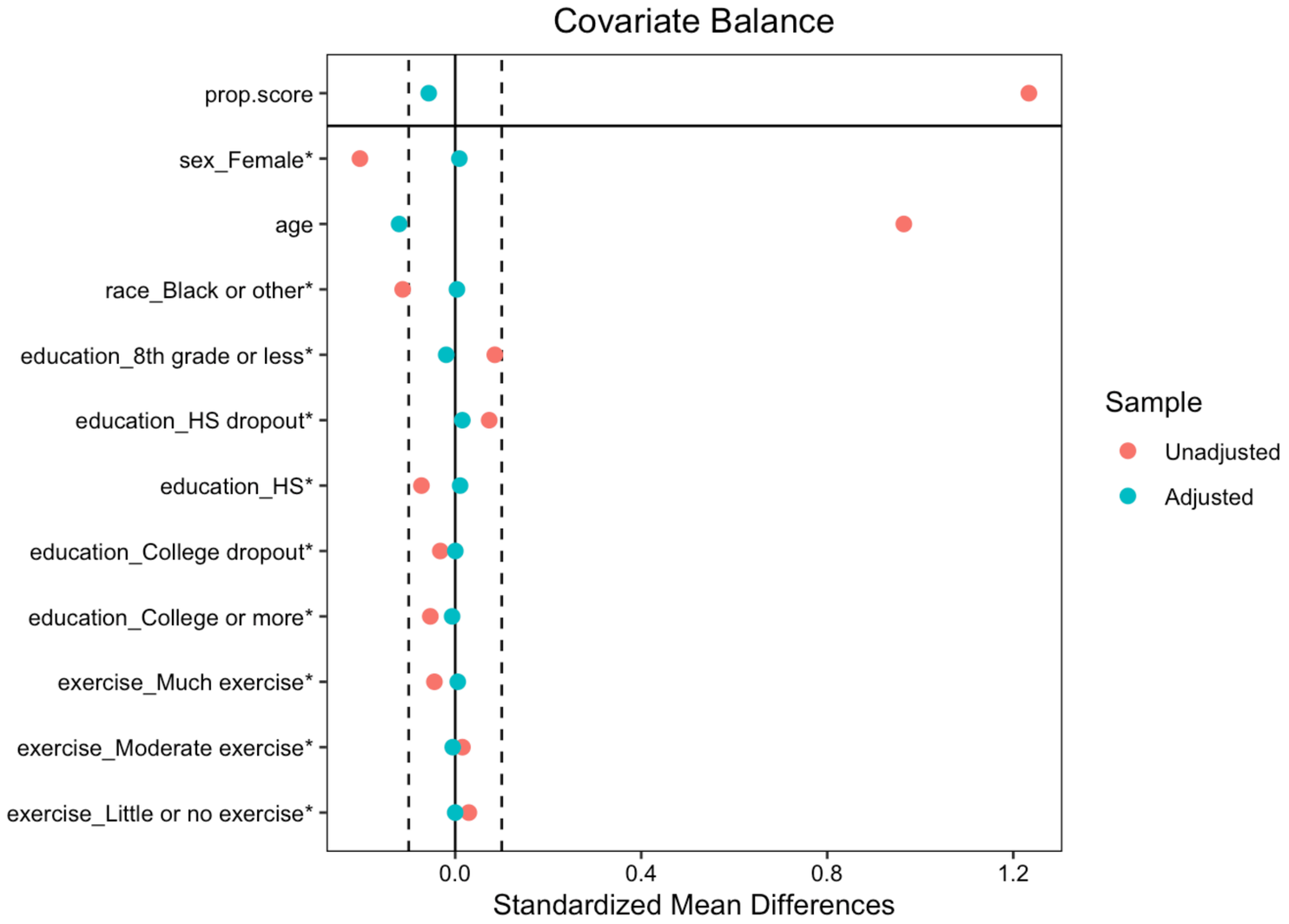
```r
# 傾向スコアモデルを作り、傾向スコアを推定、ATEを算出するための重み計算まで一気におこなう関数
ipw_model <- weightit(pack_years ~  sex + age +
                                    race + education +
                                    exercise,
                      data = df05,
                      method = "ps",
                      estimand = "ATE")

# バランスの確認
bal.plot(ipw_model,
         var.name = "prop.score",
         which = "both",
         type = "histogram",
         mirror = TRUE)
```



love plotを使ったバランスの確認（HP: 5.4.2）

- 近年は、
  群間比較によるP値ではなく、
  標準化差でバランスを確認する
  - 群ごとの平均値の差を標準偏差で割った値
- 0.1の絶対値より小さいとバランスが取れていると判断

```
love.plot(ipw_model,
          thresholds = 0.1,
          stars = "raw")
```



ATEの推定（HP: 5.4.3）

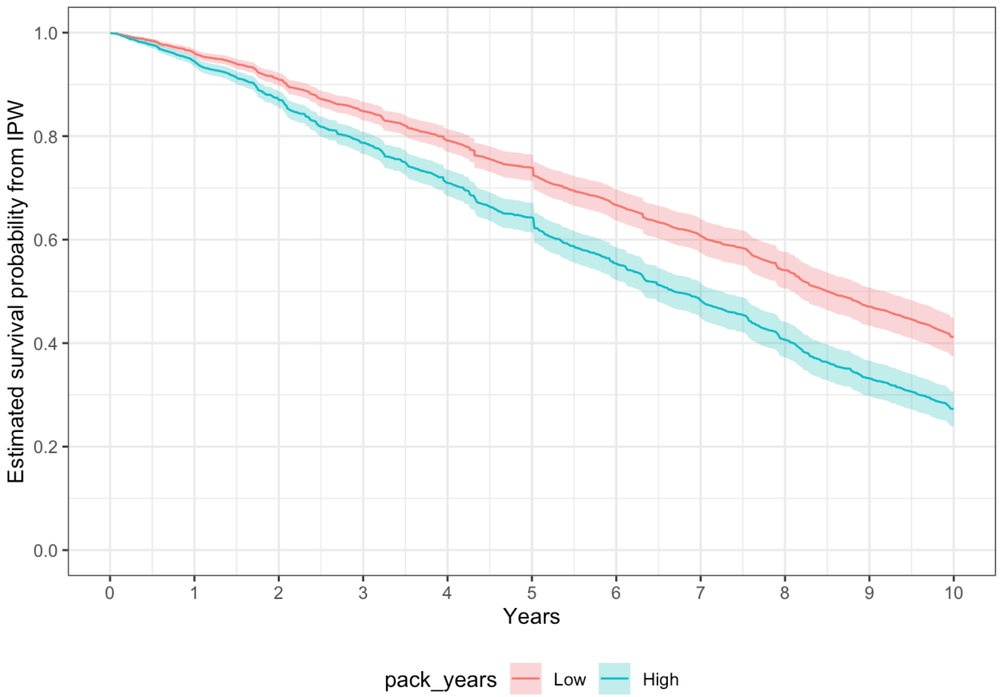
```
# 逆確率重み付け法で算出した重みを用いたCox比例ハザードモデル
cox_model_ipw <- coxph(Surv(survtime_y, death) ~ pack_years,
                       weights = ipw_model$weights,
                       cluster = seqn,
                       data = df05
                       )

# 表の作成
tab_cox_model_ipw <- tidy(cox_model_ipw, conf.int = TRUE, exponentiate = TRUE) %>%
     select(term, estimate, conf.low, conf.high, p.value)

# 表の出力
tab_cox_model_ipw |> gt()
```

term	estimate	conf.low	conf.high	p.value
pack_yearsHigh	1.464304	1.232741	1.739365	1.409734e-05

Pack-year Low群に比較してHigh群のハザード比は
1.46 (95% CI, 1.23 to 1.74)である

交絡を調整した生存確率（HP: 5.4.3）

```r
# 調整した生存確率を推定するためのデータセット作成
df06 <- df05 |>
     select(survtime_y, death, pack_years)

# 調整した生存確率を推定
estimated_survival <- predict(cox_model_ipw, df06, type = "survival", se.fit = TRUE)

# 生存確率を可視化するための準備
df07 <- bind_cols(df06,
                  estimated_survival = estimated_survival$fit,
                  ucl = estimated_survival$fit + 1.96 * estimated_survival$se.fit,
                  lcl = estimated_survival$fit - 1.96 * estimated_survival$se.fit)

# 可視化
ggplot(df07, aes(x=survtime_y, y=estimated_survival,
                 color = pack_years, fill = pack_years)) +
     geom_line() +
     geom_ribbon(aes(ymin = lcl, ymax = ucl), alpha = 0.3, color = NA) +
     xlab("Years") +
     ylab("Estimated survival probability from IPW") +
     scale_y_continuous(limits = c(0, 1),
                        breaks = seq(0, 1, by = 0.2)) +
     scale_x_continuous(limits = c(0, 10),
                        breaks = seq(0, 10, by = 1)) +
     theme_bw() +
     theme(legend.position="bottom")
```



可視化をしなくても推定はおこなえる

▼

魅力的な可視化により、
よりわかりやすく
研究のアウトプットをしよう！！

ありがとうございました！！

事後アンケートにご協力ください。