Propensity Score Diagnostics

Lucy D'Agostino McGowan

Wake Forest University

2021-09-01 (updated: 2021-10-18)

Checking balance

- Love plots (Standardized Mean Difference)
- ECDF plots

Standardized Mean Difference (SMD)

$$d = \frac{\bar{x}_{treatment} - \bar{x}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

SMD in R

```r
library(smd)
library(tidyverse)

df %>%
  # w is optional
  summarise(smd = smd(z, x, w = wts))
```

SMD in R

2 Calculate standardized mean differences

```
smds <- df %>%
  summarise(
    across(
      z1, z2, z3,
      list(
        unweighted = ~smd(.x, x)$estimate,
        weighted = ~smd(.x, x, wts)$estimate
      )
    )
  )
```

SMD in R

2 Calculate standardized mean differences

```
smds <- df %>%
  summarise(
    across(
      z1, z2, z3,
      list(
        unweighted = ~smd(.x, x)$estimate,
        weighted = ~smd(.x, x, wts)$estimate
      )
    )
  )
```

SMD in R

2 Calculate standardized mean differences

```
smds <- df %>%
  summarise(
    across(
      z1, z2, z3,
      list(
        unweighted = ~smd(.x, x)$estimate,
        weighted = ~smd(.x, x, wts)$estimate
      )
    )
  )
```

SMD in R

3 Pivot the SMDs into tidy format

```
plot_df <- smds %>%
  pivot_longer(
    everything(),
    values_to = "SMD",
    names_to = c("variable", "Method"),
    names_sep = "_"
  )
```

SMD in R

3 Pivot the SMDs into tidy format

```
plot_df <- smds %>%
  pivot_longer(
    everything(),
    values_to = "SMD",
    names_to = c("variable", "Method"),
    names_sep = "_"
  )
```

SMD in R

4 Plot them! (in a Love plot!)

```
ggplot(
  data = plot_df,
  aes(x = abs(SMD), y = variable, group = Method, color = Method)
) +
  geom_line(orientation = "y") +
  geom_point() +
  geom_vline(xintercept = 0.1, color = "black", size = 0.1)
```

SMD in R

4 Plot them! (in a Love plot!)

```
ggplot(
  data = plot_df,
  aes(x = abs(SMD), y = variable, group = Method, color = Method)
) +
  geom_line(orientation = "y") +
  geom_point() +
  geom_vline(xintercept = 0.1, color = "black", size = 0.1)
```

SMD in R

4 Plot them! (in a Love plot!)

```
ggplot(
  data = plot_df,
  aes(x = abs(SMD), y = variable, group = Method, color = Method)
) +
  geom_line(orientation = "y") +
  geom_point() +
  geom_vline(xintercept = 0.1, color = "black", size = 0.1)
```

SMD in R

4 Plot them! (in a Love plot!)

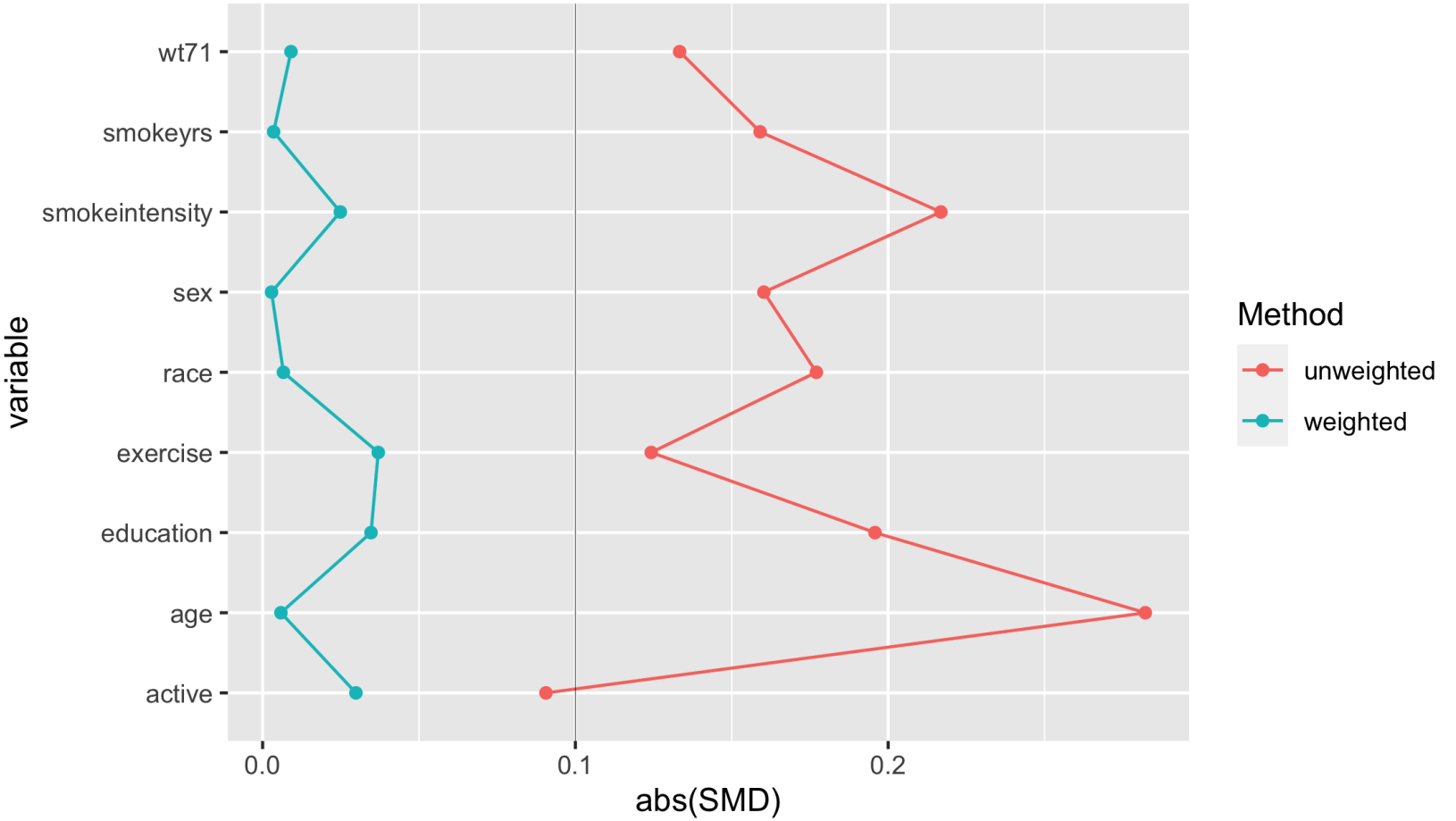
```
ggplot(
  data = plot_df,
  aes(x = abs(SMD), y = variable, group = Method, color = Method)
) +
  geom_line(orientation = "y") +
  geom_point() +
  geom_vline(xintercept = 0.1, color = "black", size = 0.1)
```

SMD in R

4 Plot them! (in a Love plot!)

```
ggplot(
  data = plot_df,
  aes(x = abs(SMD), y = variable, group = Method, color = Method)
) +
  geom_line(orientation = "y") +
  geom_point() +
  geom_vline(xintercept = 0.1, color = "black", size = 0.1)
```

Love plot


wt71
smokeyrs
smokeintensity
sex
race
exercise
education
age
active
variable
0.0
0.1
0.2
abs(SMD)
Method
unweighted
weighted

Your turn 1

1. **Create a Love Plot for the propensity score weighting you created in the previous exercise**

10:00

ECDF

For continuous variables, it can be helpful to look at the **whole** distribution pre and post-weighting rather than a single summary measure

Unweighted ECDF

```
ggplot(df, aes(x = wt71, group = qsmk, color = factor(qsmk))) +
  stat_ecdf() +
  scale_color_manual("Quit smoking", values = c("#5154B8", "#5DB854"
                     labels = c("Yes", "No")) +
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

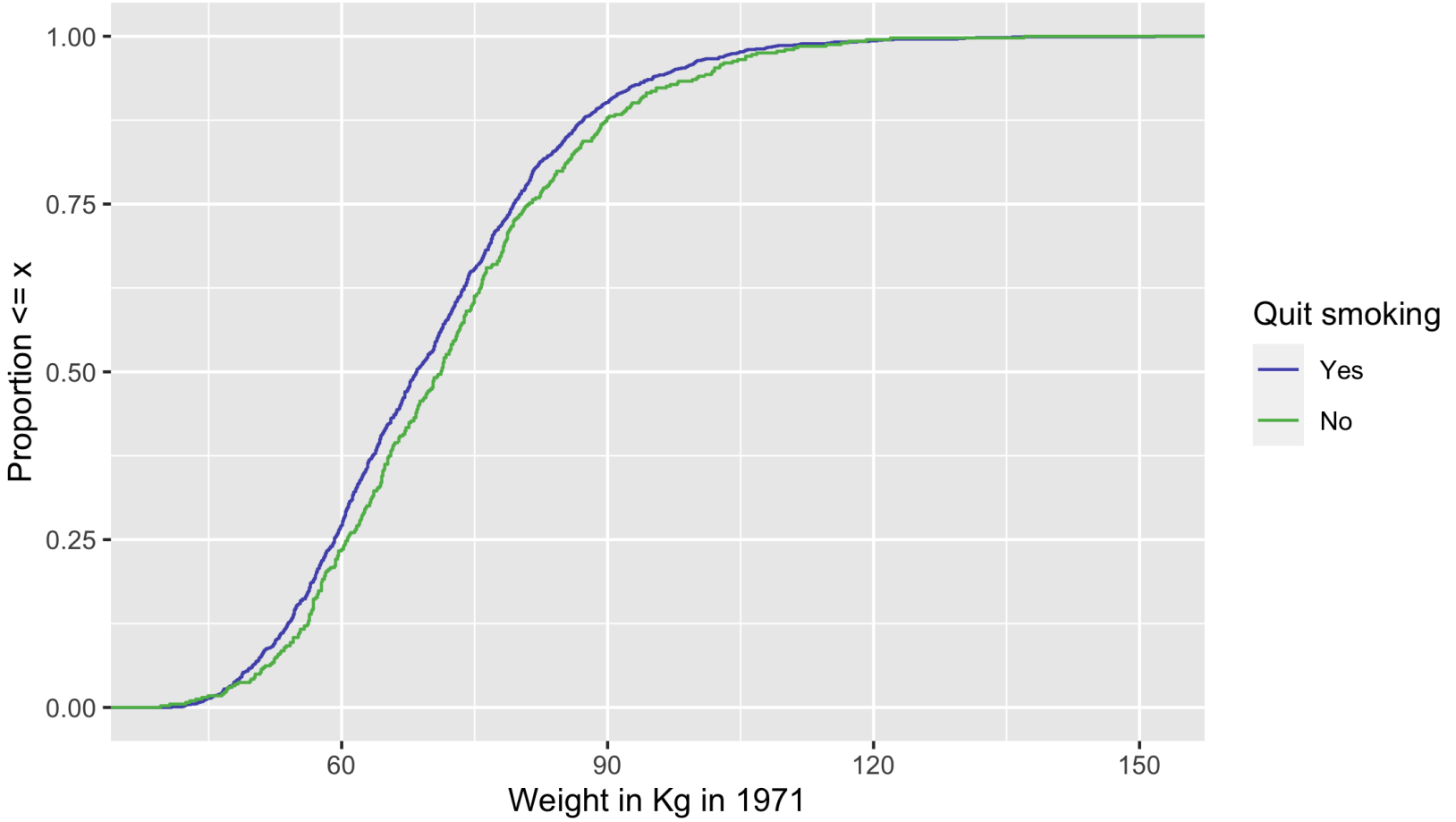
Unweighted ECDF

```
ggplot(df, aes(x = wt71, group = qsmk, color = factor(qsmk))) +
  stat_ecdf() +
  scale_color_manual("Quit smoking", values = c("#5154B8", "#5DB854"
                     labels = c("Yes", "No")) +
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

Unweighted ECDF

```
ggplot(df, aes(x = wt71, group = qsmk, color = factor(qsmk))) +
  stat_ecdf() +
  scale_color_manual("Quit smoking", values = c("#5154B8", "#5DB854"
                     labels = c("Yes", "No")) +
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

Unweighted ECDF



Weighted ECDF

```
ecdf_1 <- df %>%
  filter(qsmk == 1) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ecdf_0 <- df %>%
  filter(qsmk == 0) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ggplot(ecdf_1, aes(x = wt71, y = cum_pct)) +
  geom_line( color = "#5DB854") +
  geom_line(data = ecdf_0, aes(x = wt71, y = cum_pct), color = "#515
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

Weighted ECDF

```
ecdf_1 <- df %>%
  filter(qsmk == 1) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ecdf_0 <- df %>%
  filter(qsmk == 0) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ggplot(ecdf_1, aes(x = wt71, y = cum_pct)) +
  geom_line( color = "#5DB854") +
  geom_line(data = ecdf_0, aes(x = wt71, y = cum_pct), color = "#515
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

Weighted ECDF

```
ecdf_1 <- df %>%
  filter(qsmk == 1) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ecdf_0 <- df %>%
  filter(qsmk == 0) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ggplot(ecdf_1, aes(x = wt71, y = cum_pct)) +
  geom_line( color = "#5DB854") +
  geom_line(data = ecdf_0, aes(x = wt71, y = cum_pct), color = "#515
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

Weighted ECDF

```
ecdf_1 <- df %>%
  filter(qsmk == 1) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ecdf_0 <- df %>%
  filter(qsmk == 0) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ggplot(ecdf_1, aes(x = wt71, y = cum_pct)) +
  geom_line( color = "#5DB854") +
  geom_line(data = ecdf_0, aes(x = wt71, y = cum_pct), color = "#515
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

Weighted ECDF

```
ecdf_1 <- df %>%
  filter(qsmk == 1) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ecdf_0 <- df %>%
  filter(qsmk == 0) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ggplot(ecdf_1, aes(x = wt71, y = cum_pct)) +
  geom_line( color = "#5DB854") +
  geom_line(data = ecdf_0, aes(x = wt71, y = cum_pct), color = "#515
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

Weighted ECDF

```
ecdf_1 <- df %>%
  filter(qsmk == 1) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ecdf_0 <- df %>%
  filter(qsmk == 0) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ggplot(ecdf_1, aes(x = wt71, y = cum_pct)) +
  geom_line( color = "#5DB854") +
  geom_line(data = ecdf_0, aes(x = wt71, y = cum_pct), color = "#515
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

Weighted ECDF

```
ecdf_1 <- df %>%
  filter(qsmk == 1) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ecdf_0 <- df %>%
  filter(qsmk == 0) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ggplot(ecdf_1, aes(x = wt71, y = cum_pct)) +
  geom_line( color = "#5DB854") +
  geom_line(data = ecdf_0, aes(x = wt71, y = cum_pct), color = "#515
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

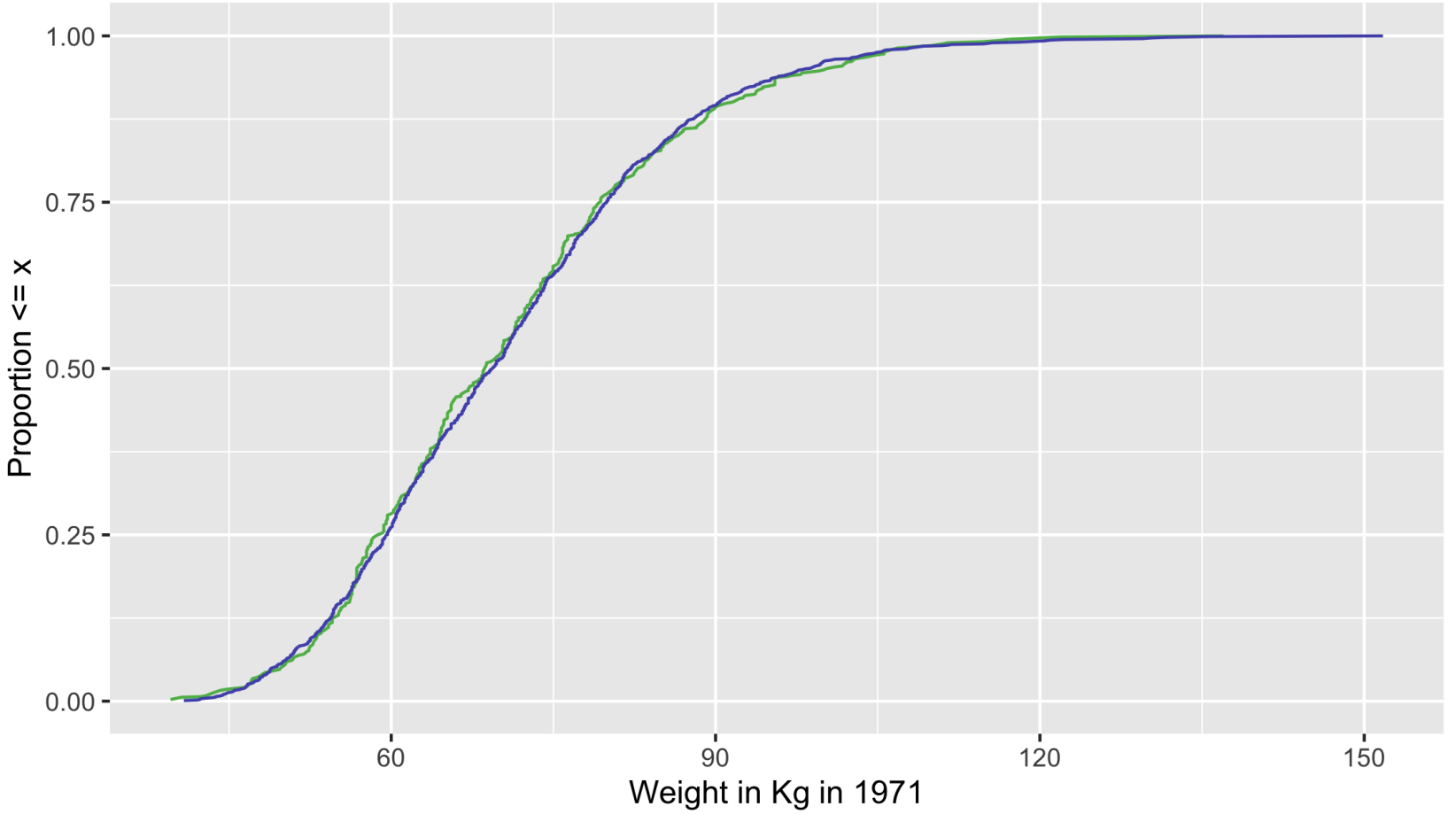
Weighted ECDF

```
ecdf_1 <- df %>%
  filter(qsmk == 1) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ecdf_0 <- df %>%
  filter(qsmk == 0) %>%
  arrange(wt71) %>%
  mutate(cum_pct = cumsum(w_ate) / sum(w_ate))

ggplot(ecdf_1, aes(x = wt71, y = cum_pct)) +
  geom_line( color = "#5DB854") +
  geom_line(data = ecdf_0, aes(x = wt71, y = cum_pct), color = "#515
  xlab("Weight in Kg in 1971") +
  ylab("Proportion <= x")
```

Weighted ECDF



Your turn 2

1. **Create an unweighted ECDF examining the** smokeyrs **confounder for those that quit smoking and those that did not**

2. **Create a weighted ECDF examining the** smokeyrs **confounder**

10:00

Bonus! Weighted Tables in R

Weighted Tables in R

1 Create a "design object" to incorporate the weights

```
library(survey)

svy_des <- svydesign(
  ids = ~ 1,
  data = df,
  weights = ~ wts,
  strata = ~ x
)
```

Weighted Tables in R

2 **Pass to** gtsummary::tbl_svysummary()

```
library(gtsummary)
tbl_svysummary(svy_des, by = x) %>%
  add_difference(everything() ~ "smd")
```

Characteristic	0, N = 1,565[1]	1, N = 1,561[1]	Difference[2]	95% CI[2,3]
WEIGHT IN KILOGRAMS IN 1971	69 (60, 80)	69 (59, 79)	0.01	-0.06, 0.08
0: WHITE 1: BLACK OR OTHER IN 1971			0.01	-0.06, 0.08
0	1,359 (87%)	1,352 (87%)		
1	206 (13%)	209 (13%)		
AGE IN 1971	43 (33, 52)	43 (33, 53)	-0.01	-0.08, 0.06
0: MALE 1: FEMALE			0.00	-0.07, 0.07
0	764 (49%)	764 (49%)		
1	802 (51%)	797 (51%)		
NUMBER OF CIGARETTES SMOKED PER DAY IN 1971	20 (10, 25)	20 (10, 30)	0.02	-0.05, 0.09
YEARS OF SMOKING	24 (15, 33)	24 (14, 33)	0.00	-0.07, 0.07
IN RECREATION, HOW MUCH EXERCISE? IN 1971, 0:much exercise,1:moderate exercise,2:little or no exercise			0.04	-0.03, 0.11
0	302 (19%)	294 (19%)		
1	665 (42%)	691 (44%)		
2	599 (38%)	576 (37%)		
IN YOUR USUAL DAY, HOW ACTIVE ARE YOU? IN 1971, 0:very active, 1:moderately active, 2:inactive			0.03	-0.04, 0.10
0	700 (45%)	684 (44%)		
1	718 (46%)	738 (47%)		
2	147 (9.4%)	138 (8.9%)		

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

[2] Standardized Mean Difference

[3] CI = Confidence Interval