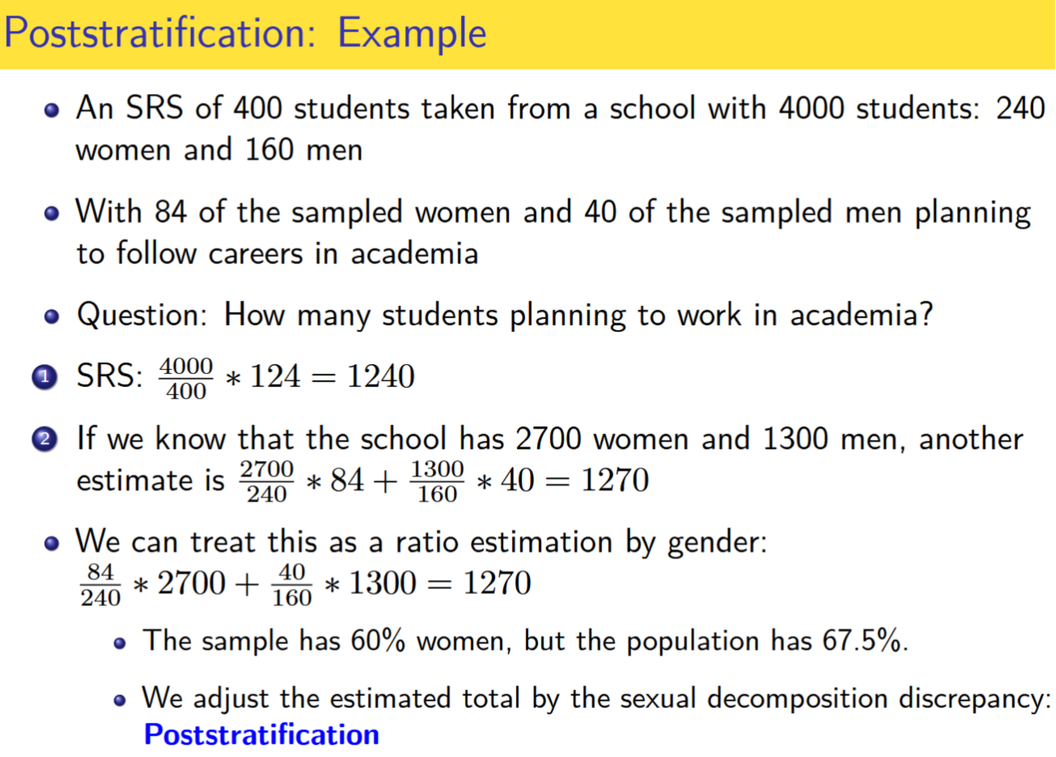
# Proposal: stratified_estimates



## Poststratification: Example

- An SRS of 400 students taken from a school with 4000 students: 240 women and 160 men
- With 84 of the sampled women and 40 of the sampled men planning to follow careers in academia
- Question: How many students planning to work in academia?

1. SRS: $\frac{4000}{400} * 124 = 1240$
2. If we know that the school has 2700 women and 1300 men, another estimate is $\frac{2700}{240} * 84 + \frac{1300}{160} * 40 = 1270$

- We can treat this as a ratio estimation by gender: $\frac{84}{240} * 2700 + \frac{40}{160} * 1300 = 1270$
  - The sample has 60% women, but the population has 67.5%.
  - We adjust the estimated total by the sexual decomposition discrepancy: **Poststratification**

### Provide proposal for point estimate by gender alongside standard errors.

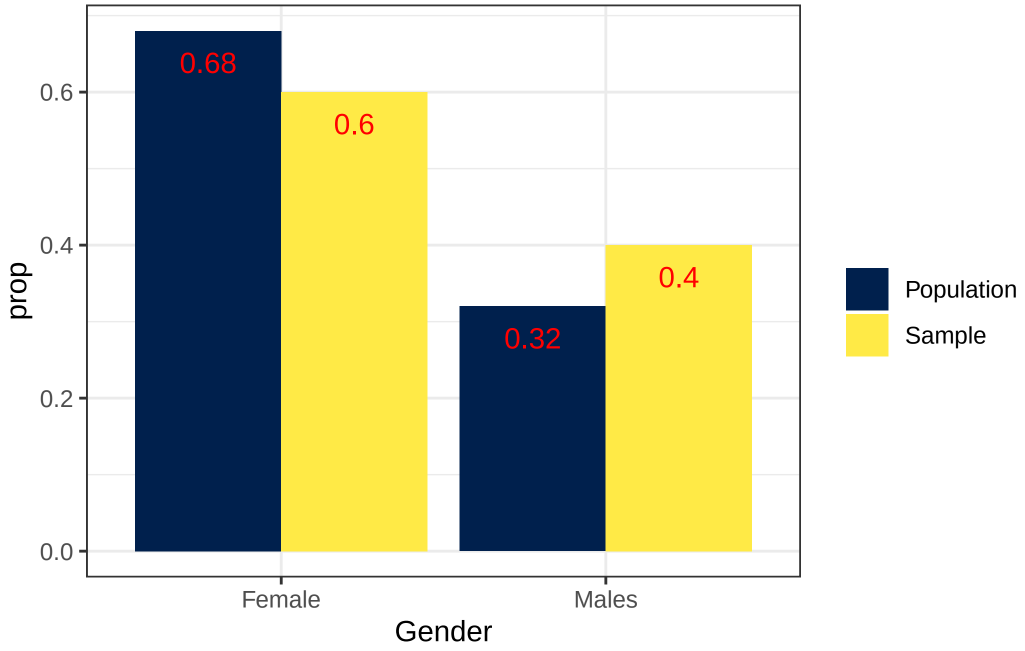
We are interested in estimating a proportion of female and male students whom are planning to pursue academic careers in a population of 4,000 (Female=2,700, Males=1300) from a sample of students n=400 (Female=240, Males=160)

```
# population
N <- 4000
N_male <- 1300
N_female <- 2700

# sample
n <- 400
n_male <- 160
n_female <- 240

# yes to academia
y_male <- 40
y_female <- 84
```

**We can explore whether our samples of females and males approximates the proportion of females and males in the population they come from.**



**We can see that we may need to up sample females, and down sample males to match our population proportions.**

## Calculating point estimates

- In our sample there are 84 females and 40 males planning to pursue academic careers. If we compute an inference for the population proportion these will still need to be calibrated to correct for sampling variability.

```
p_hat_females <- y_female/n_female ; p_hat_females
```

```
[1] 0.35
```

```
p_hat_males <- y_male/n_male ; p_hat_males
```

```
[1] 0.25
```

## Estimating variance and standard error

- We calculate the element variance estimate for a proportion for Females and Males as:

$$\hat{s}_h^2 = \frac{n_h}{n_h - 1} p_h (1 - p_h)$$

```
s_squared_female <- (n_female / (n_female - 1) ) *
  (p_hat_females *  (1 - p_hat_females)) ; s_squared_female
```

```
[1] 0.2284519
```

```
s_squared_male <- (n_male / (n_male - 1) ) *
  (p_hat_males *  (1 - p_hat_males)) ; s_squared_male
```

```
[1] 0.1886792
```

- and we can use the element variance estimate for each stratum to calculate the sampling variance estimate as;

$$var(\widehat{p}_h) = (1 - f_h)\frac{\widehat{s}_h^2}{n_h}$$

```r
var_female <- (1 - (n_female / N_female)) *
  (s_squared_female / n_female) ; var_female
```

```
[1] 0.000867271
```

```r
var_male <- (1 - (n_male / N_male)) *
  (s_squared_male / n_male) ; var_male
```

```
[1] 0.001034107
```

- which we can than use to estimate standard error by taking the square root of the variance $se(\widehat{p}_h) = \sqrt{var(\widehat{p}_h)}$.

```r
se_female <- sqrt(var_female) ; se_female
```

```
[1] 0.02944947
```

```r
se_male <- sqrt(var_male) ; se_male
```

```
[1] 0.03215754
```

### We put all estimates together in a table and estimate total count by gender alongside 95% CI.

```r
dat |>
  filter(type == "Population") |>
  select(Gender, size) |>
  add_column(
  Pnt_est = c(p_hat_females, p_hat_males),
  elem_var = c(s_squared_female, s_squared_male),
  samp_var = c(var_female, var_male),
  SE = c(se_female, se_male)
  ) |>
  mutate_if(is.double, round, 4) |>
```

```
# calculate counts
mutate(Total = Pnt_est * size,
       # calculate 95% CI to check work
       lower = round((Pnt_est - qt(.975, n-1) * SE) * size),
       upper = round((Pnt_est + qt(.975, n-1) * SE) * size)
       ) |>
knitr::kable()
```

| Gender | size | Pnt_est | elem_var | samp_var | SE | Total | lower | upper |
|---|---:|---:|---:|---:|---:|---:|---:|---:|
| Female | 2700 | 0.35 | 0.2285 | 0.0009 | 0.0294 | 945 | 789 | 1101 |
| Males | 1300 | 0.25 | 0.1887 | 0.0010 | 0.0322 | 325 | 243 | 407 |

**We propose that the population estimate from our sample to be 945 (SE=.029, [95% CI: 789, 1,101]) female and 325 (SE=.032, [95% CI: 243, 407]) male students are planing to pursue academic careers.**

**We check our work with the survey package:**

```r
# disaggregate stratified counts into dataframe
survey_data <- data.frame(
  gender = rep(c("Women", "Men"), c(n_female, n_male)),
  academia = c(rep(1, 84), rep(0, 240 - 84), rep(1, 40), rep(0, 160 - 40))
)

# created survey design object with sample and pop counts
d <- svydesign(id = ~1, data = survey_data,
               strata = ~gender,
               fpc = rep(c(N_female, N_male), c(n_female, n_male)))

# Estimate total number of students planning for academia
total_academia <- svytotal(~academia, d)

total_academia_women <- svytotal(~academia, subset(d, gender == "Women"))
# Domain estimation for men
total_academia_men <- svytotal(~academia, subset(d, gender == "Men"))
# Compute degrees of freedom for each domain
df_women <- degf(subset(d, gender == "Women"))
df_men <- degf(subset(d, gender == "Men"))
#CI
print("Point estimate for females with 95% CI = ")
```

```
[1] "Point estimate for females with 95% CI = "
```

```r
total_academia_women
```

```
         total     SE
academia   945 79.514
```

```r
confint(total_academia_women, level=.95, df=df_women)
```

```
            2.5 %   97.5 %
academia 788.3631 1101.637
```

```
print("Point estimate for males with 95% CI = ")
```

```
[1] "Point estimate for males with 95% CI = "
```

```
total_academia_men
```

```
         total     SE
academia   325 41.805
```

```
confint(total_academia_men, level=.95, df=df_men)
```

```
            2.5 %   97.5 %
academia 242.4357 407.5643
```

- We estimate the same 95% CI and point estimate for males and females. Our variance and SE is based on working with proportions instead of counts.