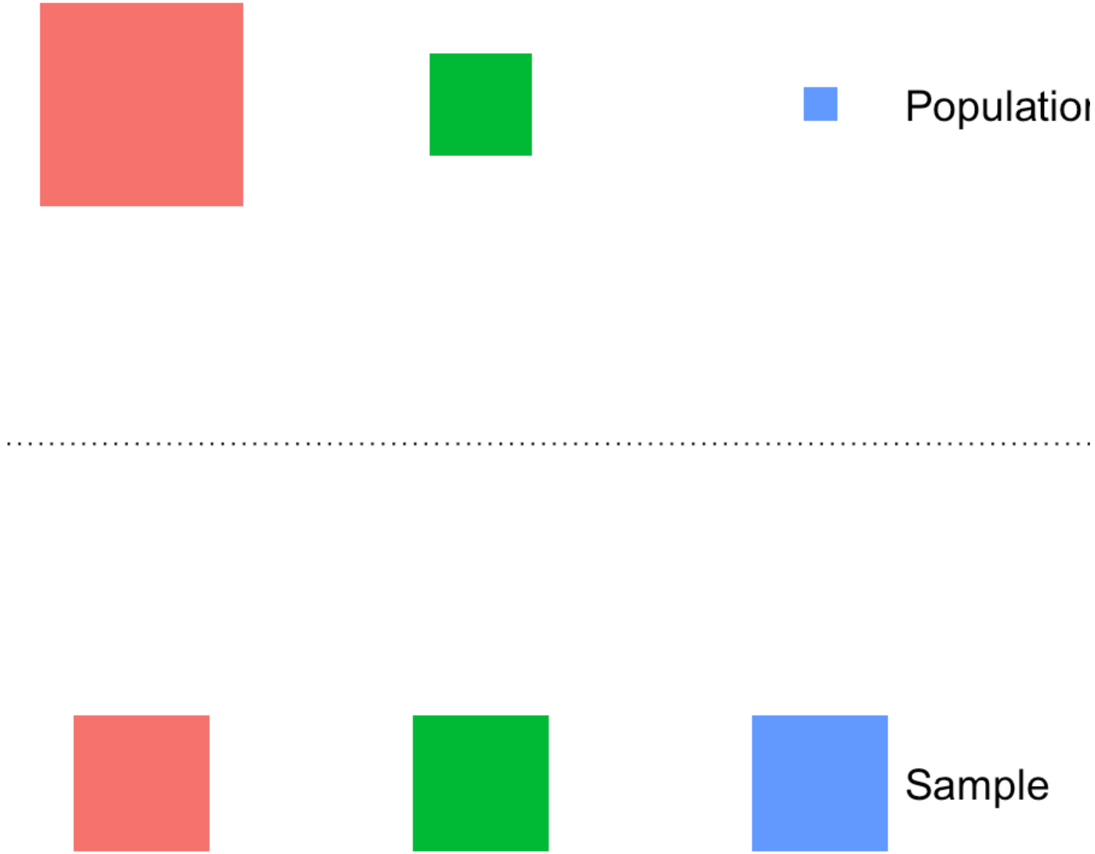
POLSCI 9590: Methods I

Sampling Weights and Final

Dave Armstrong

Sampling



Sampling Weight

A sampling weight identifies the number of people in the population for which each individual in the sample stands in.

- if π_i is the probability with which each observation is in the sample,
- $\frac{1}{\pi_i}$ is the sampling weight.

Group	Population	Sample	π_i	$\frac{1}{\pi_i}$
A	59933	500	0.0083	119.9
B	30131	500	0.0166	60.3
C	9936	500	0.0503	19.9

Estimating the Mean

R   Stata

```
library(tidyverse)
set.seed(4532)
mu <- c(-5, 0, 10)
probs=c(.6, .3, .1)
pop <- data.frame(group = sample(1:3,
                                 100000,
                                 replace=TRUE,
                                 prob = probs))
pop$y <- rnorm(100000, mu[pop$group], 2)
mean(pop$y)
```

```
## [1] -1.996288
```

```
samp <- pop %>%
  group_by(group) %>%
  mutate(n_pop = n()) %>%
  sample_n(500) %>%
  mutate(weight = n_pop/n()) %>%
  ungroup
```

- No Sampling Weights

```
samp %>% summarise(mean = mean(y))
```

```
## # A tibble: 1 × 1
##    mean
##   <dbl>
## 1  1.77
```

- With Weights

```
library(srvyr)
samp %>% as_survey_design(weights=weight) %>%
  summarise(mean = survey_mean(y))
```

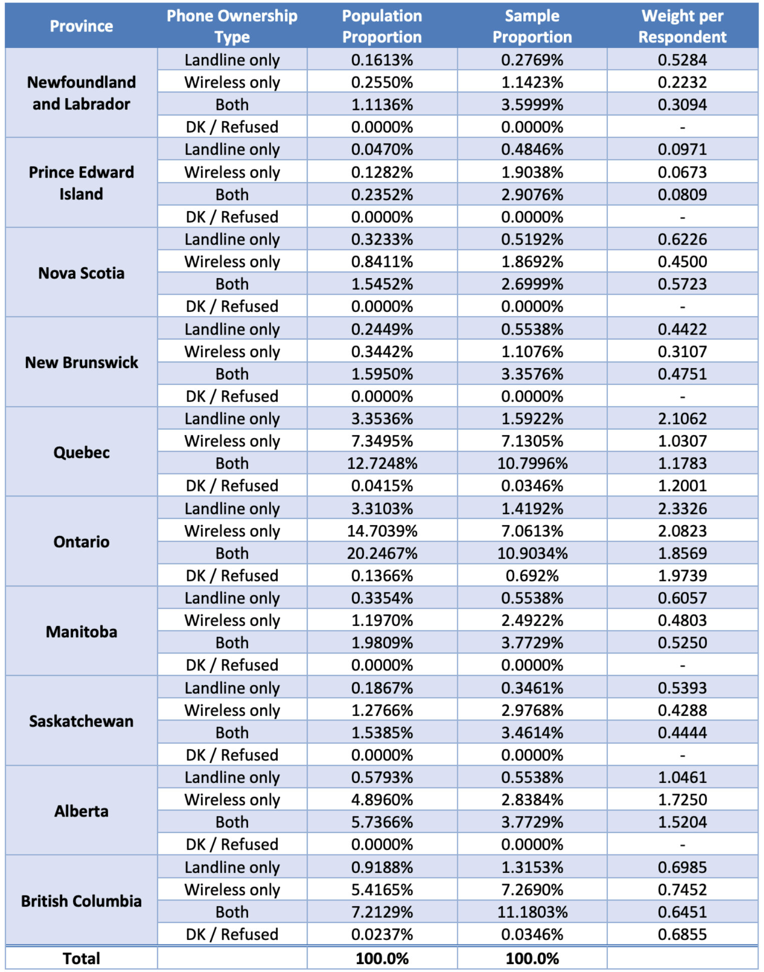
```
## # A tibble: 1 × 2
##    mean mean_se
##   <dbl>   <dbl>
## 1 -1.92   0.116
```

Weights and the CES

The CES data we've been using have weights because:

1. Each region was sampled approximately equally.
2. They made some adjustment for mobile vs landline phone usage.

Table 2.2: Weights for PES



Province	Phone Ownership Type	Population Proportion	Sample Proportion	Weight per Respondent
Newfoundland and Labrador	Landline only	0.1613%	0.2769%	0.5284
	Wireless only	0.2550%	1.1423%	0.2232
	Both	1.1136%	3.5999%	0.3094
	DK / Refused	0.0000%	0.0000%	-
Prince Edward Island	Landline only	0.0470%	0.4846%	0.0971
	Wireless only	0.1282%	1.9038%	0.0673
	Both	0.2352%	2.9076%	0.0809
	DK / Refused	0.0000%	0.0000%	-
Nova Scotia	Landline only	0.3233%	0.5192%	0.6226
	Wireless only	0.8411%	1.8692%	0.4500
	Both	1.5452%	2.6999%	0.5723
	DK / Refused	0.0000%	0.0000%	-
New Brunswick	Landline only	0.2449%	0.5538%	0.4422
	Wireless only	0.3442%	1.1076%	0.3107
	Both	1.5950%	3.3576%	0.4751
	DK / Refused	0.0000%	0.0000%	-
Quebec	Landline only	3.3536%	1.5922%	2.1062
	Wireless only	7.3495%	7.1305%	1.0307
	Both	12.7248%	10.7996%	1.1783
	DK / Refused	0.0415%	0.0346%	1.2001
Ontario	Landline only	3.3103%	1.4192%	2.3326
	Wireless only	14.7039%	7.0613%	2.0823
	Both	20.2467%	10.9034%	1.8569
	DK / Refused	0.1366%	0.692%	1.9739
Manitoba	Landline only	0.3354%	0.5538%	0.6057
	Wireless only	1.1970%	2.4922%	0.4803
	Both	1.9809%	3.7729%	0.5250
	DK / Refused	0.0000%	0.0000%	-
Saskatchewan	Landline only	0.1867%	0.3461%	0.5393
	Wireless only	1.2766%	2.9768%	0.4288
	Both	1.5385%	3.4614%	0.4444
	DK / Refused	0.0000%	0.0000%	-
Alberta	Landline only	0.5793%	0.5538%	1.0461
	Wireless only	4.8960%	2.8384%	1.7250
	Both	5.7366%	3.7729%	1.5204
	DK / Refused	0.0000%	0.0000%	-
British Columbia	Landline only	0.9188%	1.3153%	0.6985
	Wireless only	5.4165%	7.2690%	0.7452
	Both	7.2129%	11.1803%	0.6451
	DK / Refused	0.0237%	0.0346%	0.6855
Total		**100.0%**	**100.0%**	

Summaries

R   Stata

```r
library(rio)
ces19w <- import("ces19w.dta")
ces19w <- factorize(ces19w)
library(srvyr)
library(DAMisc)
cesw <- ces19w %>%
  as_survey_design(weights=weight)
sumStats(ces19w, var="market", byvar="agegrp")
```

```
## # A tibble: 3 × 12
##   variable agegrp    mean    sd   iqr    min    q25    q50   q75   max     n
##   <chr>    <fct>    <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl> <int>
## 1 market   18-34  -0.186  0.350 0.53  -0.866 -0.464 -0.2   0.066 0.734   218
## 2 market   35-54  -0.143  0.336 0.402 -1     -0.334 -0.198 0.068 0.866   498
## 3 market   55+    -0.0801 0.337 0.466 -1     -0.332 -0.066 0.134 0.866   457
## # ℹ 1 more variable: nNA <int>
```

```r
sumStats(cesw, var="market", byvar="agegrp")
```

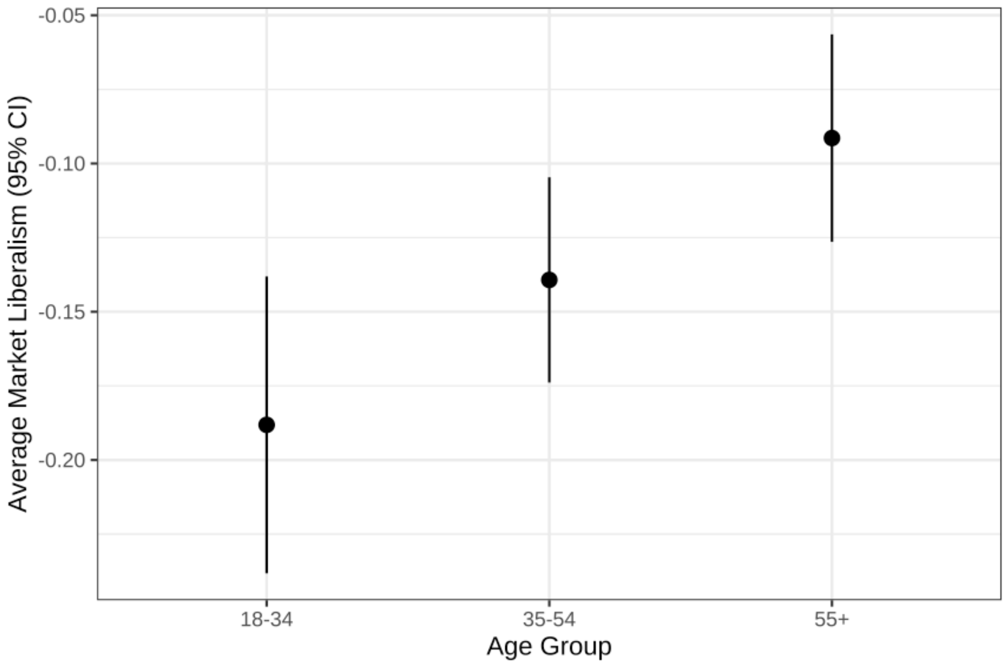
```
## # A tibble: 3 × 11
##   agegrp variable    mean    sd    min    q25 median     q75   max     n   nNA
##   <fct>  <chr>      <dbl> <dbl>  <dbl>  <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 18-34  market   -0.188  0.347 -0.866 -0.464 -0.2   0.00200 0.734  247.     0
## 2 35-54  market   -0.139  0.344 -1     -0.398 -0.198 0.068   0.866  519      0
```

Plotting Confidence Intervals

R

Stata

```
cesw %>%
 group_by(agegrp) %>%
 summarise(m = survey_mean(market, na.rm=TRUE)) %>%
 na.omit() %>%
 mutate(lwr = m-1.96*m_se,
        upr = m+1.96*m_se) %>%
 ggplot(aes(x=agegrp, y=m, ymin=lwr, ymax=upr)) +
   geom_pointrange() +
   theme_bw() +
   labs(x = "Age Group",
        y="Average Market Liberalism (95% CI)")
```



Cross-tabulations with Weights

R

Stata

```r
xt(cesw, var="vote", byvar="agegrp")
```

```
## $tab
## $tab[[1]]
##   vote/agegrp      18-34      35-54         55+          Total
##       Liberal  24%  (60)  26% (134)  31% (145)   28%   (339)
##  Conservative  30%  (75)  38% (196)  35% (165)   35%   (436)
##           NDP  22%  (54)  18%  (92)  18%  (83)   19%   (229)
##            BQ  17%  (41)  12%  (63)  11%  (49)   12%   (153)
##         Green   4%  (11)   7%  (35)   3%  (13)    5%    (59)
##         Other   2%   (6)   0%   (0)   2%  (10)    1%    (16)
##         Total 100% (247) 100% (520) 100% (465)  100% (1,232)
##
##
## $chisq
## $chisq[[1]]
##
##     Pearson's X^2: Rao & Scott adjustment
##
## data:  NextMethod()
## F = 2.499, ndf = 9.8898, ddf = 11590.8410, p-value = 0.005596
##
##
##
## $stats
```

Correlations with Weights

R Stata

```r
corfun <- function(df, var1, var2, level=.95, digits=3){
  require(survey)
  form <- glue::glue("scale({var1}) ~ scale({var2})-1")
  m = svyglm(form, design = df)
  r = coef(m)[1]
  p <- summary(m)$coef[1,4]
  r <- sprintf(glue::glue("%.{digits}f"), r)
  r <- glue::glue("{r}{ifelse(p < 1-level, '*', '')}")
  cat(glue::glue("r({var1},{var2}) = {r}\n"))
  }
corfun(cesw, "market", "leader_con")
```

```
## r(market,leader_con) = 0.291*
```

Linear Models with Weights

R Stata

```
library(survey)
w_mod <- svyglm(market ~ educ , design=cesw)
summary(w_mod)
```

```
##
## Call:
## svyglm(formula = market ~ educ, design = cesw)
##
## Survey design:
## Called via srvyr
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.08359    0.01912  -4.373 1.34e-05 ***
## educHS/College   -0.06958    0.02626  -2.649  0.00817 **
## educCollege Grad -0.06739    0.02821  -2.389  0.01706 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1171218)
##
## Number of Fisher Scoring iterations: 2
```