Tutorial 6: Stratified Sampling Solutions

Lovnesh Bhardwaj

October 21, 2024

Question (a)

What is the value of the parameter that the pollster tries to estimate?

The pollster is trying to estimate the true fraction of T voters in the entire country, denoted as $p = \frac{(24+16)}{100} = 0.4$.

Question (b)

Is the estimator unbiased?

To check whether the estimator $\hat{p_{strat1}}$ is unbiased, we need to compute its expected value and check if it is equal to the true proportion p.

The stratified estimator is given by:

$$\hat{p_{strat1}} = 0.5 \times \frac{X_r}{5} + 0.5 \times \frac{X_b}{5}$$

Where: - X_r is the number of T supporters in the Red County sample - X_b is the number of T supporters in the Blue County sample

The expected value of $\hat{p_{strat1}}$ is:

$$E[\hat{p_{strat1}}] = 0.5 \times E\left[\frac{X_r}{5}\right] + 0.5 \times E\left[\frac{X_b}{5}\right]$$

Since the expected value of the number of T supporters in the Red County is 0.8 (as 16 out of 20 support T) and in the Blue County is 0.3 (as 24 out of 80 support T), the expected value of the estimator becomes:

$$E[\hat{p_{strat1}}] = \frac{(5 \times 0.8 + 5 \times 0.3)}{10} = 0.55$$

Thus, the expected value is 0.55, which is different from the true proportion $p = 0.4$, meaning that this estimator is not unbiased.

Question (c)

Find the values of c and w to make the estimators unbiased.

For the estimator $\hat{p_{strat2a}} = c \times \hat{p_{strat1}}$, the constant c is calculated as:

$$c = \frac{0.4}{0.55} = 0.73$$

For the estimator $\hat{p_{strat2b}} = w \times \frac{X_r}{5} + (1 - w) \times \frac{X_b}{5}$, we can calculate w by solving the equation:

$$0.4 = w \times 0.8 + (1 - w) \times 0.3$$

This gives:

$$w = \frac{0.4 - 0.3}{0.5} = 0.2$$

Thus, $w = 0.2$.

Validity of the estimators:

Question (d)

Determine the standard deviation of $\hat{p_{strat2b}}$.

To calculate the variance of $\hat{p_{strat2b}}$, we start by expressing the estimator as:

$$\hat{p_{strat2b}} = w \times \frac{X_r}{5} + (1 - w) \times \frac{X_b}{5}$$

Given that $w = 0.2$, the estimator becomes:

$$\hat{p_{strat2b}} = 0.2 \times \frac{X_r}{5} + 0.8 \times \frac{X_b}{5}$$

Variance Calculation:

The variance of a weighted sum of two independent random variables is given by:

$$V(\hat{p_{strat2b}}) = V\left(0.2 \times \frac{X_r}{5} + 0.8 \times \frac{X_b}{5}\right)$$

Using the fact that $V(aX) = a^2 \times V(X)$, we expand the variance as:

$$V(\hat{p_{strat2b}}) = 0.04^2 \times V(X_r) + 0.16^2 \times V(X_b)$$

Next, we calculate the variances of X_r and X_b under a binomial distribution:

- For X_r (the number of T supporters in the Red County sample), the variance is:

$$V(X_r) = 5 \times 0.8 \times 0.2$$

- For X_b (the number of T supporters in the Blue County sample), the variance is:

$$V(X_b) = 5 \times 0.3 \times 0.7$$

Substituting these into the equation for the variance of $\hat{p_{strat2b}}$:

$$V(\hat{p_{strat2b}}) = 0.04^2 \times 5 \times 0.8 \times 0.2 + 0.16^2 \times 5 \times 0.3 \times 0.7$$

```
# Standard deviation
v <- 0.04^2 * 5 * 0.8 * 0.2 + 0.16^2 * 5 * 0.3 * 0.7
std_dev <- sqrt(v)
std_dev
```

```
## [1] 0.1678094
```

Question (f)

Voter Poll Simulation

We will simulate 10,000 repetitions of three sampling schemes:

1. **Simple Random Sampling**: We sample 10 individuals from the entire population with replacement.
2. **Proportional Stratified Random Sampling**: We sample 2 individuals from the Red County (where 80% support T) and 8 from the Blue County (where 30% support T).
3. **Weighted Stratified Random Sampling**: We sample 5 individuals from both Red County and Blue County, using the weights found in previous parts.

We will then compute the mean and standard deviation for each sampling scheme.

Simulation Code:

```
# Set the number of simulations
n <- 1000000

# Simple Random Sampling: sample 10 individuals from the population with p = 0.4 (T s
upporters)
p.srs <- rbinom(n, 10, 0.4) / 10

# Proportional Stratified Random Sampling: sample 2 from Red County (p = 0.8) and 8 f
rom Blue County (p = 0.3)
p.str <- (rbinom(n, 2, 0.8) + rbinom(n, 8, 0.3)) / 10

# Weighted Stratified Random Sampling: sample 5 from Red County (p = 0.8) and 5 from
Blue County (p = 0.3)
p.str2b <- 0.2 * rbinom(n, 5, 0.8) / 5 + 0.8 * rbinom(n, 5, 0.3) / 5

# Calculate the mean and standard deviation for each sampling method
mean_srs <- mean(p.srs)
mean_str <- mean(p.str)
mean_str2b <- mean(p.str2b)

sd_srs <- sd(p.srs)
sd_str <- sd(p.str)
sd_str2b <- sd(p.str2b)

# Output the results
mean_srs
```

```
## [1] 0.4001189
```

```
mean_str
```

```
## [1] 0.4000938
```

```
mean_str2b
```

```
## [1] 0.4000144
```

```
sd_srs
```

```
## [1] 0.1550635
```

```
sd_str
```

```
## [1] 0.1415909
```

```
sd_str2b
```

```
## [1] 0.1677687
```

Question (g)

What is the approximate distribution of $\hat{p_{strat2b}}$?

Based on the Central Limit Theorem, we know that the sum of a large number of independent random variables tends to follow a normal distribution. In this case, $\hat{p_{strat2b}}$, being a weighted sum of two independent binomially distributed random variables (representing samples from Red and Blue counties), will approximately follow a normal distribution as the number of samples becomes large.

Therefore, the approximate distribution of $\hat{p_{strat2b}}$ is:

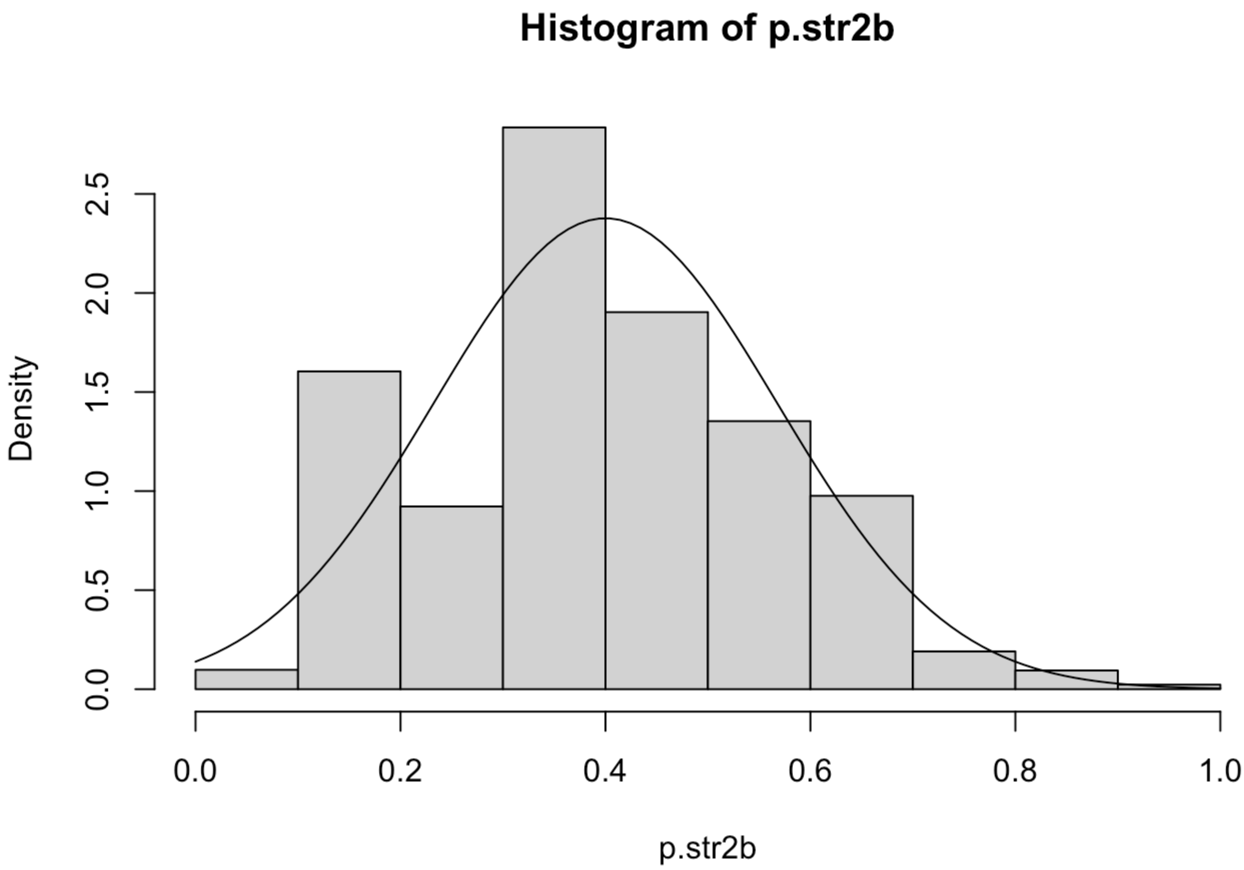
$$\hat{p_{strat2b}} \sim N(0.4, 0.17^2)$$

This means that $\hat{p_{strat2b}}$ is normally distributed with a mean of 0.4 and a standard deviation of 0.17. The mean reflects the true proportion of T supporters, and the standard deviation indicates the spread or variability in the estimator.

Question (h)

Make a histogram of the values for $\hat{p_{strat2b}}$ and superimpose the density of the approximate distribution.

```
# Histogram with density plot
hist(p.str2b,breaks=10,prob=TRUE)
x<-seq(0,1,length=100)
lines(x,dnorm(x,0.4,sqrt(v)))
```


Histogram of p.str2b
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
p.str2b