Assignment 3a

Satvik Saha

2024-09-17

Answer 1

(a) We approximate the 95% confidence interval for the proportion p who answer 'Yes' using a $\pm z_{0.025}$ s.e. interval around the estimate $\hat{p} = y/n$.

```
n <- 1500
y <- 750

p.hat <- y / n
se    <- sqrt(p.hat * (1 - p.hat) / n)
p.ci  <- p.hat + qnorm(c(0.025, 0.975)) * se

p.ci
```

```
## [1] 0.474697 0.525303
```

(b)

```
n <- 1500
y <- 900

p.hat <- y / n
se    <- sqrt(p.hat * (1 - p.hat) / n)
p.ci  <- p.hat + qnorm(c(0.025, 0.975)) * se

p.ci
```

```
## [1] 0.5752082 0.6247918
```

(c)

```
n <- 10
y <- 6

p.hat <- y / n
se    <- sqrt(p.hat * (1 - p.hat) / n)
p.ci  <- p.hat + qnorm(c(0.025, 0.975)) * se

p.ci
```

```
## [1] 0.2963637 0.9036363
```

Answer 2

We switch to using an estimator of the form $\hat{p} = (y + 2)/(n + 4)$.

```r
n <- 50
y <- 0

p.hat <- (y + 2) / (n + 4)
se    <- sqrt(p.hat * (1 - p.hat) / n)
p.ci  <- p.hat + qnorm(c(0.025, 0.975)) * se

p.ci
```

```
## [1] -0.01530926  0.08938334
```

Thus, we report a 95% confidence interval of $[0, 0.09]$.

Answer 3

Let $\ell = 0.15$ be the national fraction of Latinos. Also let p_L be the probability of a 'Yes' response from a Latino, p_O from others. Further assume that $p_L, p_O \approx 0.5$. We model the two options for the survey as follows.

(i) Let each response be recorded as (x_i, z_i), where x_i is 1 for 'Yes' and 0 for 'No', and z_i is 1 for Latino and 0 for other. Then, we have

$$x_i \mid z_i \sim \text{Bernoulli}(p_L^{z_i} p_O^{1-z_i}), \qquad z_i \sim \text{Bernoulli}(\ell).$$

It follows that

$$x_i \sim \text{Bernoulli}(p), \qquad p = \ell p_L + (1-\ell) p_O.$$

Thus, p denotes the national average. Setting $y = \sum_i x_i$, the number of 'Yes' responses from a simple random sample of $n = 1000$ Americans is

$$y \sim \text{Binomial}(n, p).$$

The number of responses from Latinos is $n_L = \sum_i z_i$, others is $n_O = \sum_i (1 - z_i) = n - n_L$. Also, the number of 'Yes' responses from Latinos is $y_L = \sum_i x_i z_i$, others is $y_O = \sum_i x_i (1 - z_i) = y - y_L$. We note that y_L, y_O when conditioned on n_L are independent; this model for $(n_L, n_O, y_L, y_O)$is equivalent to first drawing $n_L \sim \text{Binomial}(n, \ell)$, setting $n_O = n - n_L$, then drawing (independently)

$$y_L \mid n_L \sim \text{Binomial}(n_L, p_L), \qquad y_O \mid n_L \sim \text{Binomial}(n_O, p_O).$$

With this, the difference $p_L - p_O$ between Latinos and others is estimated by $y_L/n_L - y_O/n_O$, and the national average p is estimated as y/n.

(ii) The number of Latinos sampled is $n_L = 300$, non-Latinos is $n_O = 700$, for a total of $n = n_L + n_O = 1000$ responses. The number of 'Yes' responses y_L from Latinos and y_O from others is modeled as

$$y_L \sim \text{Binomial}(n_L, p_L), \qquad y_O \sim \text{Binomial}(n_O, p_O).$$

The difference $p_L - p_O$ between Latinos and others is estimated as $y_L/n_L - y_O/n_O$, and the national average p is estimated as $\ell y_L/n_L + (1-\ell) y_O/n_O$.

With this, we can now compare the two options.

(a) Option (ii) gives more accurate comparisons between Latinos and others, and option (i) gives more accurate national population estimates.

(b) First, examine the national average estimates. In option (i), we simply have $\text{var}(y/n) = p(1-p)/n \approx 1/4n$. In option (ii), we have

$$\text{var}\left(\frac{\ell y_L}{n_L} + \frac{(1-\ell) y_O}{n_O}\right) = \frac{\ell^2 p_L (1-p_L)}{n_L} + \frac{(1-\ell)^2 p_O (1-p_O)}{n_O} \approx \frac{\ell^2}{4n_L} + \frac{(1-\ell)^2}{4n_O}.$$

```r
n   <- 1000
n_L <- 300
n_O <- 700
ell <- 0.15

print(sqrt(1/(4 * n)))
```

```
## [1] 0.01581139
```

```r
print(sqrt(1/4 * (ell^2 / n_L + (1 - ell)^2 / n_O)))
```

```
## [1] 0.01663688
```

Thus, option (i) gives a tighter estimate of the national average opinion.

Next, examine the comparison between the Latino response and the others. In option (i), observe that

$$\text{var}\left(\frac{y_L}{n_L} - \frac{y_O}{n_O}\right) = \mathbb{E}\left[\text{var}\left(\frac{y_L}{n_L} - \frac{y_O}{n_O}\,\middle|\, n_L\right)\right] + \text{var}\left(\mathbb{E}\left[\frac{y_L}{n_L} - \frac{y_O}{n_O}\,\middle|\, n_L\right]\right).$$

Note that the second term vanishes; the conditional expectation there is just the constant $p_L - p_O$. This leaves us with

$$\mathbb{E}\left[\frac{p_L(1-p_L)}{n_L} + \frac{p_O(1-p_O)}{n_O}\right] \approx \mathbb{E}\left[\frac{1}{4n_L} + \frac{1}{4(n-n_L)}\right] = \mathbb{E}\left[\frac{n}{4n_L(n-n_L)}\right].$$

Recall that $n_L \sim \text{Binomial}(n, \ell)$; we will approximate the above via simulation.

In option (ii), we simply have

$$\text{var}\left(\frac{y_L}{n_L} - \frac{y_O}{n_O}\right) = \frac{p_L(1-p_L)}{n_L} + \frac{p_O(1-p_O)}{n_O} \approx \frac{1}{4n_L} + \frac{1}{4n_O} = \frac{n}{4n_L(1-n_L)}.$$

```r
n   <- 1000
n_L <- 300
n_O <- 700
ell <- 0.15

n_L.sim <- rbinom(1000000, 1000, ell)

print(sqrt(mean(n/(4 * n_L.sim * (n - n_L.sim)))))
```

```
## [1] 0.04438997
```

```r
print(sqrt(1/4 * (1 / n_L + 1 / n_O)))
```

```
## [1] 0.03450328
```

Thus, option (ii) gives a tighter estimate for the difference in response between Latinos and others.

Research homework assignment

We generate the data $\theta_j \sim t_5(0, 0.1)$ and $\hat{\theta}_j \mid \theta_j \sim N(\theta_j, \sigma_j)$ for $1 \le j \le J$ as follows.

```r
generate.theta <- function(
        J,
        sigma,
        df = 5,
        scale = 0.1
```

```
) {
        theta <- rt(J, df = df) * scale
        theta.hat <- rnorm(J, mean = theta, sd = sigma)
        return(list(
                theta = theta,
                theta.hat = theta.hat
        ))
}
```

With this, we construct the suggested strategies, referred to as "Everything", "Nothing", "Positive", and "Significant" (in order). Each strategy takes a J-vector of $(\hat{\theta}_j)$ and returns a binary J-vector indicating whether or not to implement each intervention.

```
strategy.everything <- function(theta.hat, sigma) {
        return(rep(1, length(theta.hat)))
}

strategy.nothing <- function(theta.hat, sigma) {
        return(rep(0, length(theta.hat)))
}

strategy.positive <- function(theta.hat, sigma) {
        return(as.numeric(theta.hat > 0))
}

z.alpha <- qnorm(0.95)
strategy.significant <- function(theta.hat, sigma) {
        return(as.numeric(theta.hat / sigma > z.alpha))
}
```

The gain for a strategy $\hat{\theta} \mapsto \delta(\hat{\theta})$ is simply $\sum_j x_j \theta_j \delta(\hat{\theta}_j)$. The ideal strategy is one that will implement an intervention *if and only if* $\theta_j > 0$. Observe that the "Nothing" strategy will always give a gain of 0; with this in mind, we will omit this strategy in further discussion.

```
strategies <- c(
        strategy.everything,
        # strategy.nothing,
        strategy.positive,
        strategy.significant
)
strategies.names <- c(
        "Everything",
        # "Nothing",
        "Positive",
        "Significant"
)
```

We loop our simulation 1000 times, for $J = 10, 10^2, 10^3, 10^4$ and for $\sigma = 0.05, 0.2$.

```
x <- 1000
gains <- data.frame()
runs <- 1000
for (J in c(10, 100, 1000, 10000)) {
        for (sigma in c(0.05, 0.2)) {
                for (i in 1:runs) {
                        d <- generate.theta(J, sigma = sigma)
```

```
                        g <- c(J, sigma)
                        for (strategy in strategies) {
                                gain <- sum(x * d$theta * strategy(d$theta.hat, sigma))
                                g <- c(g, gain)
                        }
                        gains <- rbind(gains, g)
                }
        }
}
colnames(gains) <- c("J", "sigma", strategies.names)
```

The expected gain in sales for each of these strategies can be estimated by the mean gain over the 1000 runs, as follows.

```
suppressMessages(library(tidyr))
suppressMessages(library(dplyr))

gains <- gather(gains, key = "Strategy", value = "Gains", -J, -sigma)
gains %>%
        group_by(sigma, J, Strategy) %>%
        summarize(Mean = mean(Gains)) %>%
        ungroup() %>%
        pivot_wider(names_from = Strategy, values_from = Mean)
```

```
## `summarise()` has grouped output by 'sigma', 'J'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 8 x 5
##   sigma     J Everything Positive Significant
##   <dbl> <dbl>      <dbl>    <dbl>       <dbl>
## 1  0.05    10       6.32     437.        359.
## 2  0.05   100     -28.5     4329.       3553.
## 3  0.05  1000     119.     43537.      35676.
## 4  0.05 10000    -221.    435221.     356896.
## 5  0.2     10      -7.54     252.        106.
## 6  0.2    100     -87.0     2538.       1099.
## 7  0.2   1000      17.0    25580.      11004.
## 8  0.2  10000    -205.    255418.     109453.
```

The variation in gain in sales over multiple runs can be better visualized using the following histograms.

```
library(ggplot2)

pdf("3a_rha.pdf", width = 16, height = 8)
ggplot(gains, aes(Gains)) +
        geom_histogram(
            aes(color = Strategy, fill = Strategy),
            alpha = 0.2,
            position = "identity"
        ) +
        facet_grid(rows = vars(sigma), cols = vars(J), scales = "free_x")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
dev.off()
```

It seems that on average, the "Positive" strategy performs better than the "Significant" strategy, which in

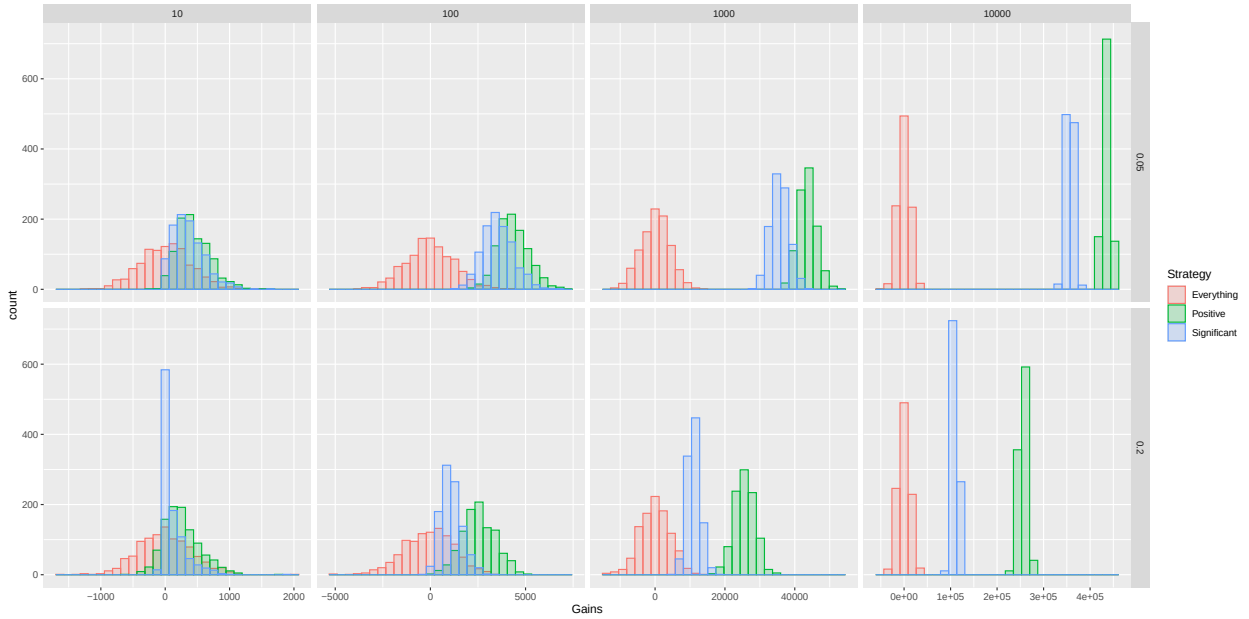


Figure 1: Distributions of gain in sales by strategy (color), J (columns), σ (rows).

turn performs better than the "Everything" and "Nothing" strategies. Both the "Everything" and "Nothing" strategies both must have an expected gain of zero, the "Everything" strategy simply has more variance. The difference between the "Positive" and "Significant" strategies becomes more apparent with increasing J.

The distribution of gains $\sum_j x_j \theta_j \delta(\hat{\theta}_j)$ for a strategy δ is asymptotically normal via the Central Limit Theorem; note that the $\theta_j \delta(\hat{\theta}_j)$ are independent and identically distributed, with finite second moments.

Note that the "Significant" strategy δ_s is more demanding than the "Positive" strategy δ_p, in the sense that $\delta_s \leq \delta_p$. By being more conservative, the "Significant" strategy seems to avoid losses (when $\theta_j < 0$), but also discards potentially small gains (when $\theta_j > 0$ is small). This may help explain why this performs poorer than the "Positive" strategy.

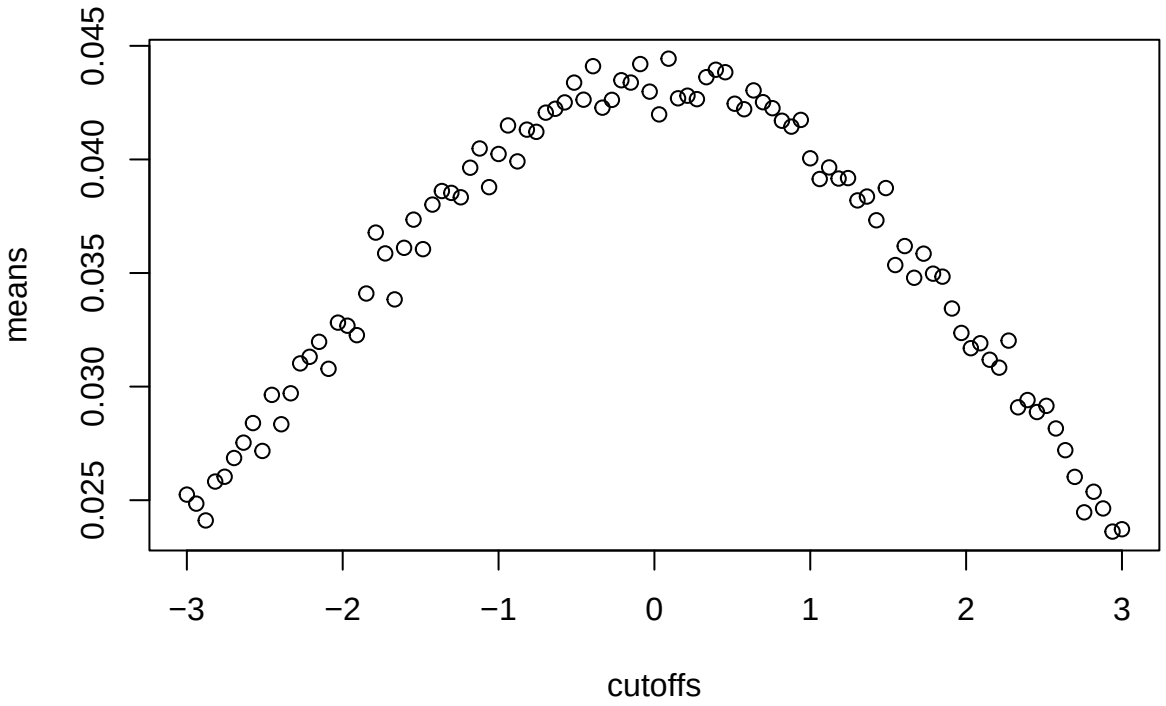
By increasing σ_j, the observed quantity $\hat{\theta}_j$ begins offering less information about θ. Thus, the "Positive" and "Significant" strategies begin suffering in performance.

In general, a strategy of the form $\delta_c(\hat{\theta}) = \mathbf{1}(\hat{\theta}/\sigma > c)$ for some cutoff c will have an expected gain of the form

$$\mathbb{E}[\theta\, \delta_c(\hat{\theta})] = \mathbb{E}[\theta\, \mathbb{P}(\hat{\theta} > c\sigma \mid \theta)] = \mathbb{E}\left[\theta\, \Phi\left(\frac{\theta}{\sigma} - c\right)\right],$$

which is difficult to calculate analytically (but has been estimated here via simulation). Of course, this quantity goes to 0 as $c \to \infty$ (DCT).

```
sigma   <- 0.05
cutoffs <- seq(-3, 3, length.out = 100)
means   <- rowMeans(replicate(10000, {
  theta <- rt(100, df = 5) * 0.1
  theta * pnorm(theta / sigma - cutoffs)
}))
plot(cutoffs, means)
```



It seems that $c = 0$ is indeed an optimal cutoff.