Assignment 3b

Satvik Saha

2024-09-23

Answer 1

(a) The number of shots made is simply a Binomial$(10, 0.40)$ random variable.

```
rbinom(1, 10, 0.40)
```

```
## [1] 5
```

(b) Each shot behaves like a Bernoulli(p_i) random variable, with $p_i = i/10$.

```
p <- seq(0.10, 1.00, by = 0.10)
sum(sapply(p, function(p.i) rbinom(1, 1, p.i)))
```

```
## [1] 5
```

Answer 2

Using the arguments from Assignment 3a, Problem 3, we take a simple random sample of n people, of which $n_m \sim \text{Binomial}(n, f_m)$ will be men, where $f_m \approx 0.5$ is the proportion of men in voting population. Then, the number of men y_m and women y_w supporting the candidate are modeled as

$$y_m \mid n_m \sim \text{Binomial}(n_m, p_m), \qquad y_w \mid n_m \sim \text{Binomial}(n_w, p_w),$$

where $n_w = n - n_m$, and p_m, p_w are the population proportions of men and women supporting the candidate. We estimate the gender gap $\delta = p_m - p_w$ via

$$\hat{\delta} = \begin{cases} y_m/n_m - y_w/n_w, & \text{if } n_m, n_w > 0, \\ 0, & \text{otherwise.} \end{cases}$$

This estimator suffers for $n_m, n_w = 0$; we set these aside as very low probability events. As before, we can compute

$$\begin{aligned}
\text{var}(\hat{\delta}) &= \mathbb{E}\left[\left(\frac{p_m(1-p_m)}{n_m} + \frac{p_w(1-p_w)}{n_w}\right)\mathbf{1}(n_m \notin \{0, n\})\right] + \text{var}\left((p_m - p_w)\mathbf{1}(n_m \notin \{0, n\})\right) \\
&\leq \mathbb{E}\left[\frac{n}{4n_m(n-n_m)}\mathbf{1}(n_m \notin \{0, n\})\right] + (p_m - p_w)^2(f_m^n + f_w^n)(1 - (f_m^n + f_w^n)) \\
&\leq \mathbb{E}\left[\frac{n}{4n_m(n-n_m)}\mathbf{1}(n_m \notin \{0, n\})\right] + (f_m^n + f_w^n) \\
&\approx \mathbb{E}\left[\frac{n}{4n_m(n-n_m)}\mathbf{1}(n_m \notin \{0, n\})\right],
\end{aligned}$$

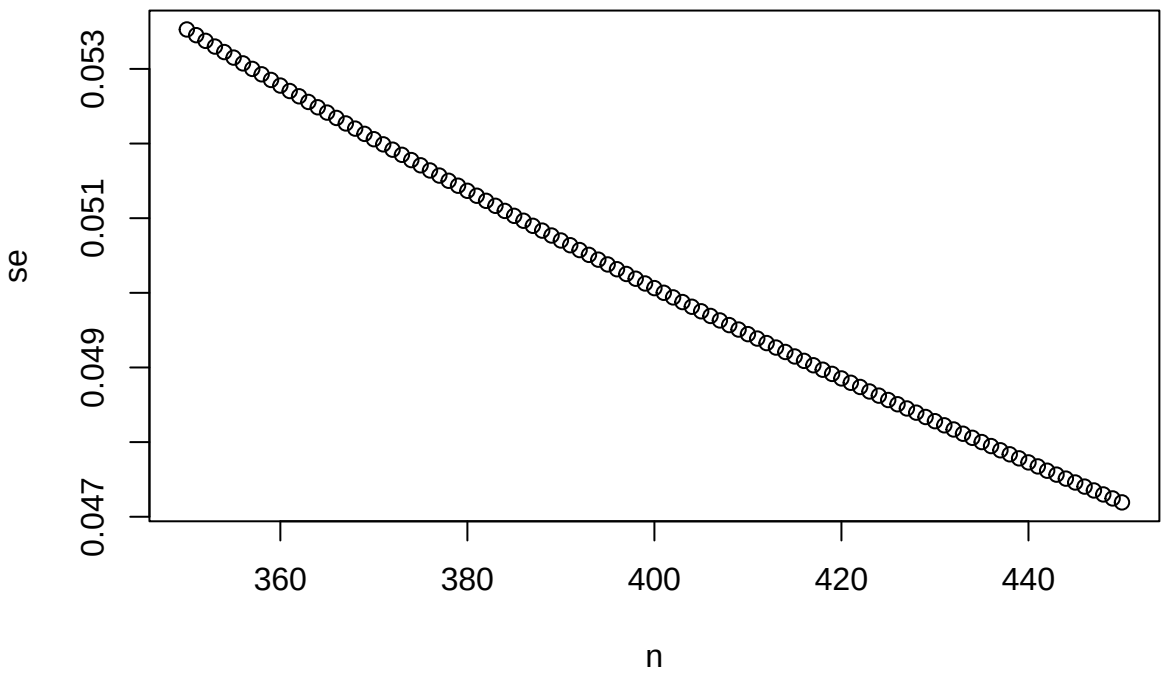
where we discard the last term $f_m^n + f_w^n \approx 1/2^{n-1}$. We simulate

```
sd.approx <- function(n, f_m = 0.5, sims = 10000) {
  n_m <- rbinom(sims, n, f_m)
```

```
  sqrt(mean(n / (4 * n_m * (n - n_m))))
}
n  <- 350:450
se <- sapply(n, function(k) sd.approx(k))
print(n[which.max(se <= 0.05)])
```

```
## [1] 402
```

```
plot(n, se)
```



Thus, we need to poll a little over 400 people to ensure that the standard error is less than 5 percentage points.

Answer 3

Let $X \sim \text{Binomial}(20, 0.30)$ and $Y \sim \text{Binomial}(20, 0.40)$ be independent, denoting the number of shots made by the two students. We have

$$\mathbb{P}(X = i, Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j) = \binom{20}{i}\binom{20}{j} 0.3^i 0.7^{20-i} 0.4^j 0.6^{20-j}.$$

Then,

$$\mathbb{P}(X < Y) = \sum_{i=0}^{20} \sum_{j=i+1}^{20} \mathbb{P}(X = i, Y = j).$$

This can be computed as approximately 69.3%, as follows.

```
probs.greater <- sapply(0:20, function(i)
  sapply((i + 1):20, function(j)
    dbinom(i, 20, 0.30) * dbinom(j, 20, 0.40)
  )
)
sum(unlist(probs.greater))
```

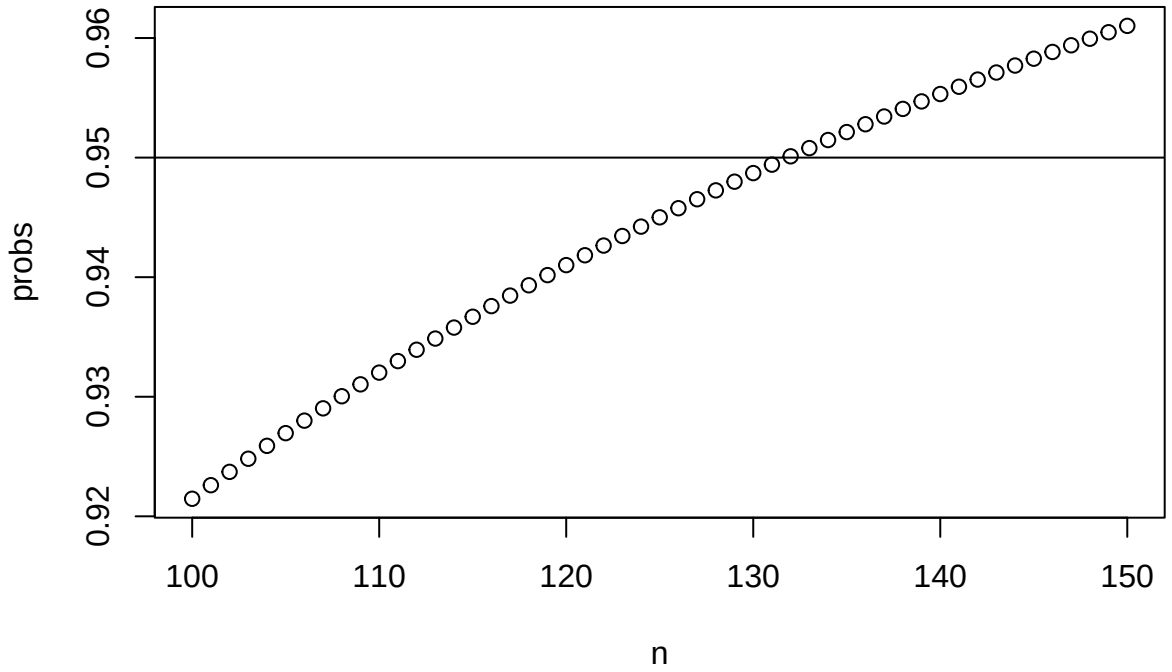
```
## [1] 0.6926396
```

Answer 4

We will say that we have a good chance of distinguishing the shooters given n shots each if the probability of the better shooter scoring higher is at least $1 - \alpha$. We choose $\alpha = 0.05$.

```
p.distinguish <- function(n, p1, p2) {
  probs.greater <- sapply(0:n, function(i)
  sapply((i + 1):n, function(j)
    dbinom(i, n, p1) * dbinom(j, n, p2)
  )
)
  sum(unlist(probs.greater))
}

n <- 100:150
probs <- sapply(n, function(k) p.distinguish(k, 0.30, 0.40))
plot(n, probs)
abline(h = 0.95)
```



With this, we demand around $n = 140$ shots.

We could also have tried using the normal approximations for independent $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(n, q)$ to obtain

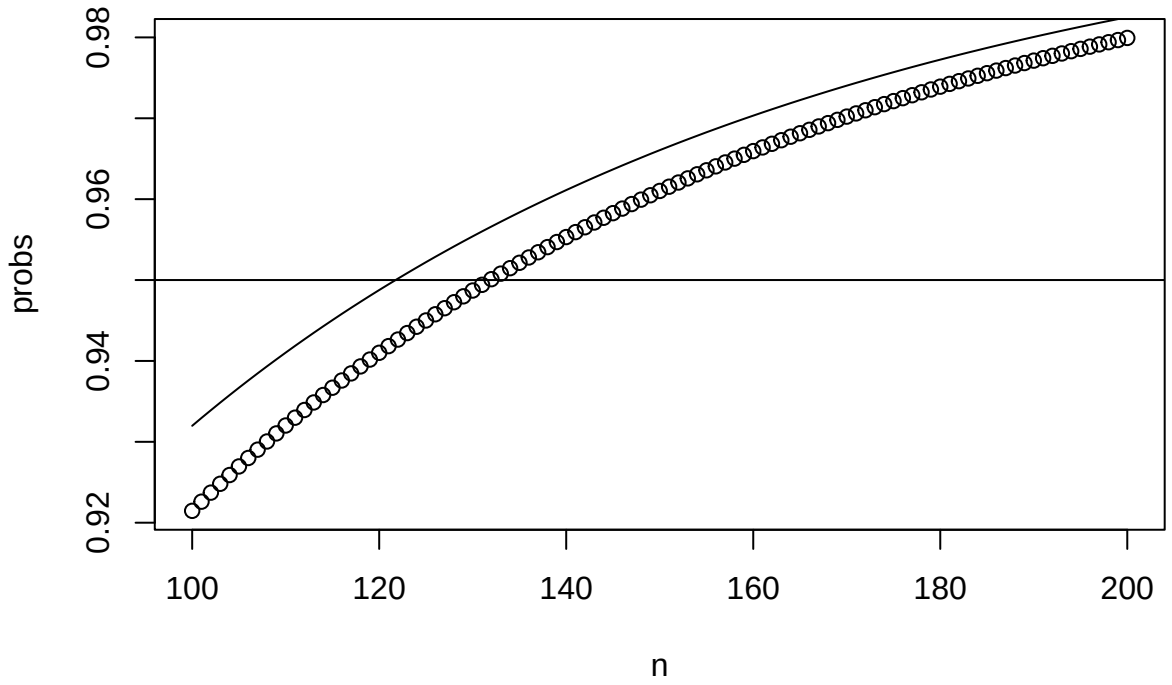
$$X - Y \overset{a}{\sim} N\left(n(p-q),\, n(p(1-p) + q(1-q))\right),$$

so

$$P(X < Y) \approx \Phi\left(\frac{\sqrt{n}(q-p)}{\sqrt{p(1-p) + q(1-q)}}\right).$$

```
p.distinguish.normal <- function(n, p1, p2) {
  pnorm(sqrt(n) * (p2 - p1) / sqrt(p1 * (1 - p1) + p2 * (1 - p2)))
}
n <- 100:200
probs <- sapply(n, function(k) p.distinguish(k, 0.30, 0.40))
probs.normal <- sapply(n, function(k) p.distinguish.normal(k, 0.30, 0.40))
```

```r
plot(n, probs)
lines(n, probs.normal)
abline(h = 0.95)
```



It seems that the normal approximation underestimates n a little.

We could also verify our answer $n \approx 140$ via simulation.

```r
mean(rbinom(10000, 140, 0.3) < rbinom(10000, 140, 0.4))
```

```
## [1] 0.9575
```