Introduction to Inference

Adapted by Juan Pablo Lewinger from OpenIntro Statistics by Diez et al.

6/20/2024

Estimating a population proportion

Example: 1,000 people poll shows US President approval rating is about 39%

- 39% is an ***estimate*** of the true but ***unknown*** approval rating among the entire adult US population, which is the parameter of interest.
- Unless we ask and get answers from every adult in the US (close to impossible), the true approval rating will remain unknown.
- We denote the true unknown approval rating parameter by p and its estimate (39% = 0.39) by $\widehat{p}$
- The error due to sampling is the difference between the parameter and its estimate: $p - \widehat{p}$
- If we took a different sample we'll get a different estimate and a different error
- What can we say in general about the error when estimating a proportion?

Poll simulation

- Assume the true approval rating is 41% ($p = 0.41$)
- Simulate US adult population of 258,000,000
- Simulate a poll: draw a random sample of 1,000 individuals
- Calculate the proportion in the sample

Poll simulation

```
pop_size <- 258000000; poll_size = 1000

p <- 0.41


USpop <- c(rep("Approve", p*pop_size), rep("Disapprove", (1-p)*pop_size))

length(USpop)
```

```
## [1] 258000000
```

```
head(USpop); tail(USpop)
```

```
## [1] "Approve" "Approve" "Approve" "Approve" "Approve" "Approve"
```

```
## [1] "Disapprove" "Disapprove" "Disapprove" "Disapprove" "Disapprove"
## [6] "Disapprove"
```

Poll simulation

```r
set.seed(2024)

poll <- sample(USpop, poll_size)

head(poll)
```

```
## [1] "Approve"    "Disapprove" "Approve"    "Disapprove" "Approve"   
## [6] "Disapprove"
```

```r
table(poll)
```

```
## poll
##    Approve Disapprove 
##        403        597
```

```r
p_hat <- sum(poll=="Approve")/poll_size

p_hat
```

```
## [1] 0.403
```

Poll simulation

If we took a different sample we'd get a different estimate:

```
poll2 <- sample(USpop, poll_size)

p_hat2 <- sum(poll2=="Approve")/poll_size

p_hat2
```

```
## [1] 0.407
```

```
poll3 <- sample(USpop, poll_size)

p_hat3 <- sum(poll3=="Approve")/poll_size

p_hat3
```

```
## [1] 0.422
```

Poll simulation

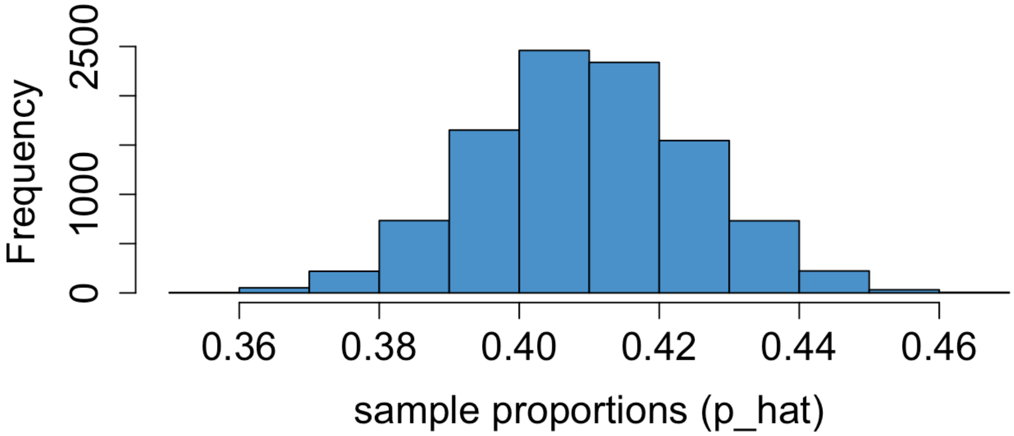
- To get a sense of the distribution of possible values of $\widehat{p}$ let's repeat the simulation many times, say 10,000, and plot a histogram of the values of $\widehat{p}$ that we get:

```
sample_proportions <- replicate(10000, sum(sample(USpop, poll_size) == "Approve")/poll_size)

head(sample_proportions)
```

```
## [1] 0.422 0.418 0.416 0.392 0.419 0.414
```

Poll simulation

```
hist(sample_proportions , col='steelblue3', xlab="sample proportions (p_hat)", main="", cex.lab=1.5, cex.axis=1.5)
```



```
c(mean = mean(sample_proportions), sd = sd(sample_proportions))
```

```
##   mean     sd
## 0.4102 0.0155
```

Poll simulation

- This is called the sampling distribution of the estimate $\widehat{p}$
- The mean of the distribution, $\mu_{\widehat{p}}$, is 0.41, the same as the true population parameter!
- This means that on average the population proportion estimates the true proportion without bias
- The spread (sd), called the standard error of $\widehat{p}$, and denoted $SE_{\widehat{p}}$ is quite small (0.016). This is the average error we make when we estimate the parameter p by its estimate $\widehat{p}$.
- When the true proportion is $p = 0.41$ and the sample size $n = 1,000$ the sample proportion tends to give a very good estimate of the population proportion
- The sampling distribution is symmetric and bell shaped, it looks like a normal distribution.

VERY IMPORTANT: The sampling distribution is never observed in real applications because we take a single sample of size n. Here we are simulating what would happen if we hypothetically took many many samples of size n.

Poll simulation

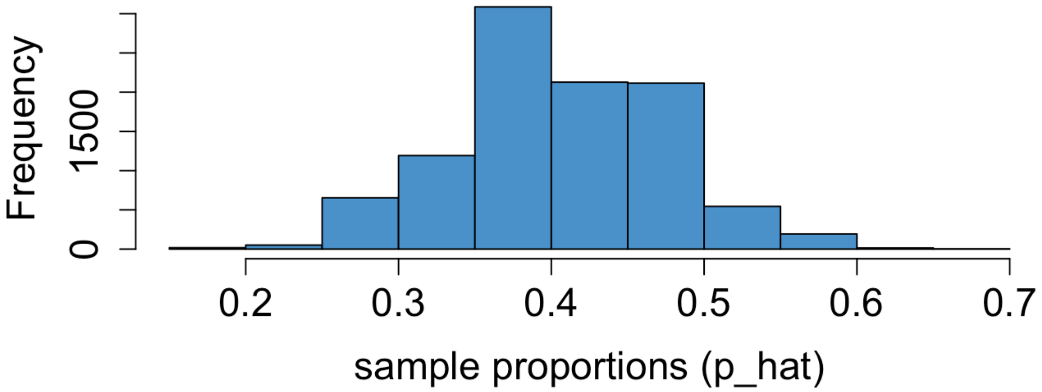
- What would the sampling distribution look like if the size of the sample was $n = 50$ instead of $n = 1,000$?

```
poll_size = 50

sample_proportions <- replicate(10000, sum(sample(USpop, poll_size) == "Approve")/poll_size)

head(sample_proportions)
```

```
## [1] 0.40 0.40 0.42 0.32 0.40 0.30
```

Poll simulation

```r
hist(sample_proportions , col='steelblue3', xlab="sample proportions (p_hat)", main="", cex.lab=1.5, cex.axis=1.5)
```



```r
c(mean = mean(sample_proportions), sd = sd(sample_proportions))
```

```
##   mean     sd
## 0.4107 0.0691
```

Poll simulation

- The mean of the sampling distribution ($\mu_{\hat{p}}$) again equals the true population parameter $p = 0.39$

- However, the standard error, $SE_{\hat{p}}$, is now about 4 times larger! (0.07 vs 0.016)

- The larger the sample size the smaller the error (the error decreases proportionally to $\frac{1}{\sqrt{n}}$

Central Limit Theorem (CLT)

When the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with the following mean and standard deviation:

$$\mu_{\hat{p}} = p$$

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when:

$$np \geq 10$$

and

$$n(1-p) \geq 10$$

This is called the success-failure condition.

Central Limit Theorem (CLT)

In our scenario $p = 0.41$ so:

$np = 1000 \times 0.41 = 410 > 10$

and

$n(1 - p) = 1000 \times 0.590 = 590 > 10$

So the Central Limit Theorem holds with:

$\mu_{\hat{p}} = 0.41$

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.41\times(1-0.41)}{1000}} = 0.0156$$

These CLT-based calculations of the mean of the sample distribution and the standard error are consistent with our simulation results.

The CLT allows us to theoretically derive the standard error of sample distributions (no need to simulate)!

Confidence interval for a proportion

- When we estimate a proportion we can report the 'point estimate' $\hat{p}$ along its standard error, which quantifies the uncertainty about the estimate.
- A point estimate will never 'hit' the true parameter exactly
- So, we would like to provide a range of values, an interval, that contains the true parameter with high confidence
- We build the confidence interval around the most plausible value, the sample proportion
- When the central limit applies, the sampling distribution is close to a normal distribution
- And normal distribution always has 95% of the data within 1.96 standard deviations of the mean
- So we construct a 95% confidence interval as:

$$Point\,Estimate \pm 1.96 \times SE$$

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence interval for a proportion

In the presidential approval example (first poll):

$\widehat{p} = 0.39$

$$\widehat{SE}_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \sqrt{\frac{0.39\times(1-0.39)}{1000}} = 0.0154$$

95% CI: $\widehat{p} \pm 1.96 \times SE_{\widehat{p}} = 0.39 \pm 1.96 \times 0.0154 = (0.360, 0.420)$

If the sample size was $n = 50$ instead:

$$\widehat{SE}_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \sqrt{\frac{0.39\times(1-0.39)}{50}} = 0.0690$$

95% CI: $\widehat{p} \pm 1.96 \times SE_{\widehat{p}} = 0.39 \pm 1.96 \times 0.069 = (0.255, 0.525)$

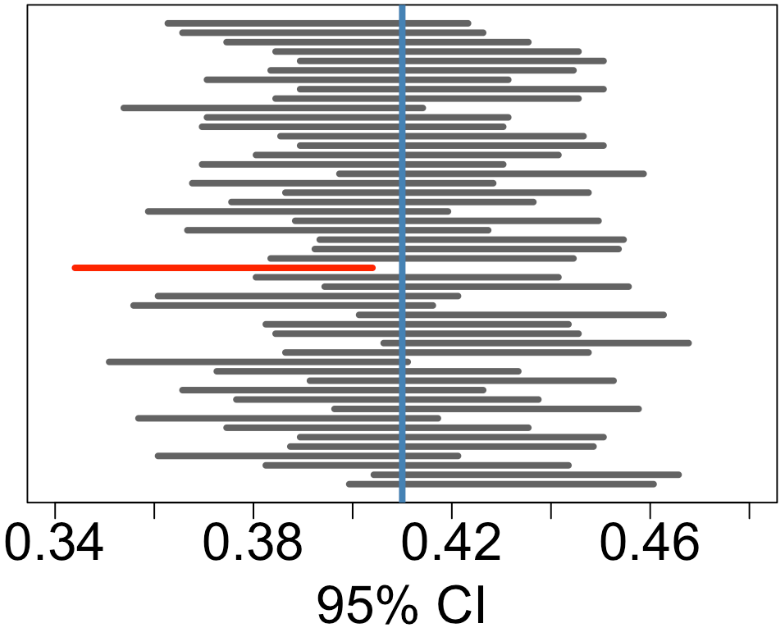
Interpretation of a 95% confidence interval

Suppose we took many samples of size n and built a 95% confidence interval from each. Then about 95% of those intervals would contain the parameter p

VERY IMPORTANT In a real application/analysis we draw a single sample of size n, compute a single point estimate, and a single confidence interval based on the sample. Here, using theory (CLT) and simulations, we are exploring what would happen if we hypothetically repeated the process of drawing a sample and computed estimates and CIs many times.

Confidence interval for a proportion

50 95% confidence intervals based on 50 different samples of size $n = 1,000$



Estimating a population mean

- We estimate the population mean μ by the population sample $\bar{x}$
- The CLT also applies to the sample mean (in the vast majority of cases):

When the sample size is sufficiently large, the sample mean $\bar{x}$ will tend to follow a normal distribution with the following mean and standard deviation:

$$\mu_{\bar{x}} = \mu$$

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation

Confidence interval for the mean

$$Point\ Estimate \pm 1.96 \times SE$$

$$\bar{x} \pm 1.96 \times \frac{\widehat{\sigma}}{\sqrt{n}}$$

where $\widehat{\sigma}$ is the sample standard deviation, which estimates the population standard deviation σ