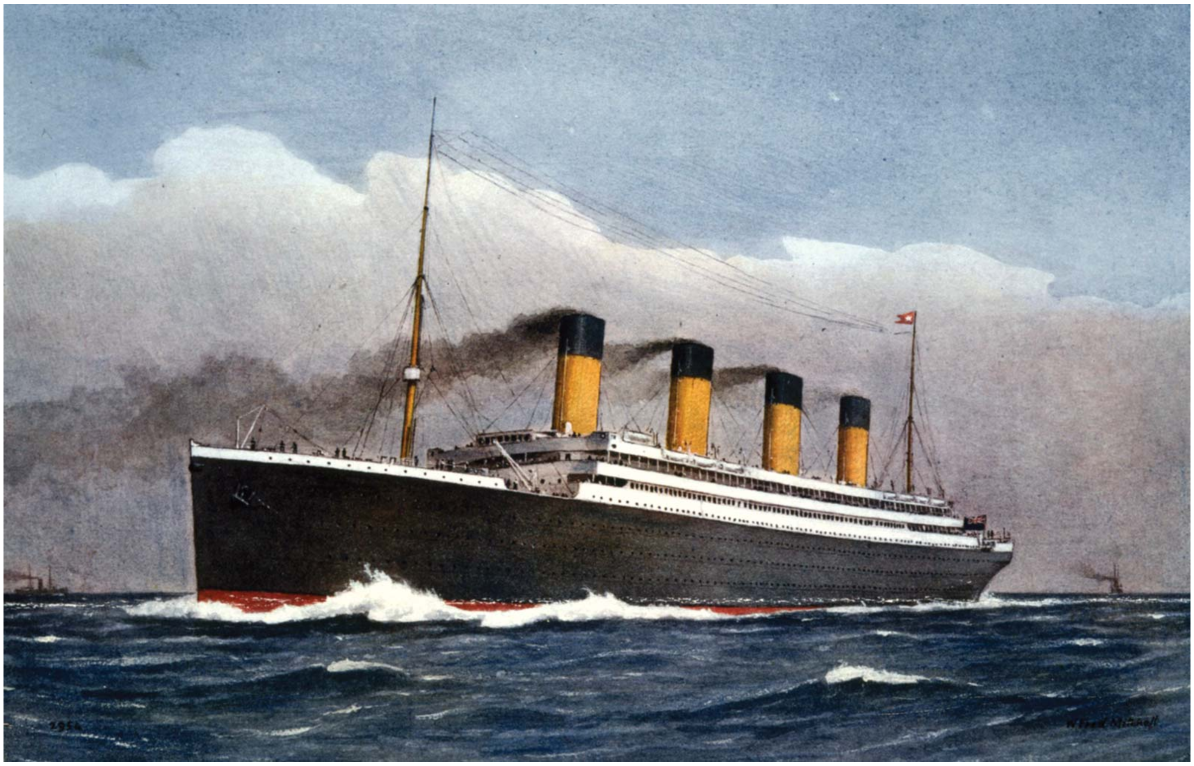
Lecture 3.1 - Practicing Confidence Intervals

Student

2025-11-13



Interpretation steps

Let's practice with a dataset of the passengers of the Titanic, from the dataset `titanic_survival.csv`.

Setting things up

1. Create new variable called `survivedindicator`; assign it a value of 0 if the passenger did not survive and 1 if the passenger did survive using the `case_when` verb (information on how to do that can be found here or here) . Find the proportion that survived in the entire dataset (you can do this by using the `table()` command: `table(titanic$survivedindicator)`.

```
titanic.pax <- titanic.pax %>%
  mutate(survivedindicator = case_when(survived == "yes" ~ 1,
                                       survived == "no" ~ 0))

titanic.pax %>%
  group_by(survivedindicator) %>%
  summarise(Count = n(),
            Percent = round(100 * n() / nrow(titanic.pax), 1)) %>%
  select(survivedindicator, Count, Percent) %>%
  kable(caption = "Summary of Passenger Survival",
        col.names = c("Survived", "Count", "Percent (%)"),
        align = "lcc") %>%
  kable_styling()
```

Table 1: Summary of Passenger Survival

Survived	Count	Percent (%)
0	619	59.2
1	427	40.8

Random matters

Now let's sample from the dataset. We can sample whether or not they survived by running the following code (note: you need the library `dplyr` loaded for this code):

```
titanic_n50 <- titanic_survival %>%
  slice_sample(n=50)
```

Create a sample of size 50 and 200.

```
tp.50 <- titanic.pax %>%
  slice_sample(n=50)

tp.200 <- titanic.pax %>%
  slice_sample(n=200)
```

2. Calculate by hand the standard error and 95% confidence interval for `survivedindicator` of your sample of 50 and 200.

   For $n = 50$, the calculation is:

- $0.36 \pm z_{95}\sqrt{\frac{\hat{0.36}(1-\hat{0.36})}{50}}$
- $0.36 \pm 1.96 \times 0.07$
- 0.36 ± 0.1372
- $(0.2228, 0.4972)$

   For $n = 200$, the calculation is:

- $0.44 \pm z_{95}\sqrt{\frac{\hat{0.44}(1-\hat{0.44})}{200}}$
- $0.44 \pm 1.96 \times 0.04$
- 0.44 ± 0.0784
- $(0.3616, 0.5184)$

3. Interpret these confidence intervals

   We are 95% confident that our confidence interval contains the true proportion. In practical terms, the actual proportion of percentage survived seems likely to be less than half.

4. Find the proportion that survived of the entire dataset – was it inside or outside the standard error of your confidence intervals for 50 and 200? Why was it inside or outside?

The true p is: 0.41

Confidence interval for the sample of 50 DID contain the true p

Confidence interval for the sample of 200 DID contain the true p

5. If you sampled many times, how many sample proportions would be inside or outside of your confidence interval you just created?

   By definition, 95%.

Sampling distributions

Let's now add the following command to the `setup` block (copy and paste the entire part and then run your `setup` code block again:

```
prop.multiple.samples <- function(n, numsamples, variable) {
     meanvector <- c()
     meanonesample <- 0
     for (i in 1:numsamples) {
       meanonesample <- mean(sample(variable, n, replace=TRUE))
         meanvector[i] <- meanonesample
     }
     meanvector
}
```

This defines a new function in R called `prop.multiple.samples()`. It takes as its arguments the sample size (`n`), the number of samples (`numsamples`) and the variable from which you would like to create a sampling distribution. Once you have created this function, you can use it as follows:

```
titanic_n50_s100 <-prop.multiple.samples(50, 100, titanic_survival$survivedindicator)
```

This takes 100 samples of $n = 50$ from the variable `titanic_survival$survivedindicator`. In practical terms, this line of code draws 50 people at random from the dataset 100 times and then calculate the proportion in each sample of the number of people surviving.

This process creates a pseudo sampling distribution.

Observing the sampling distribution

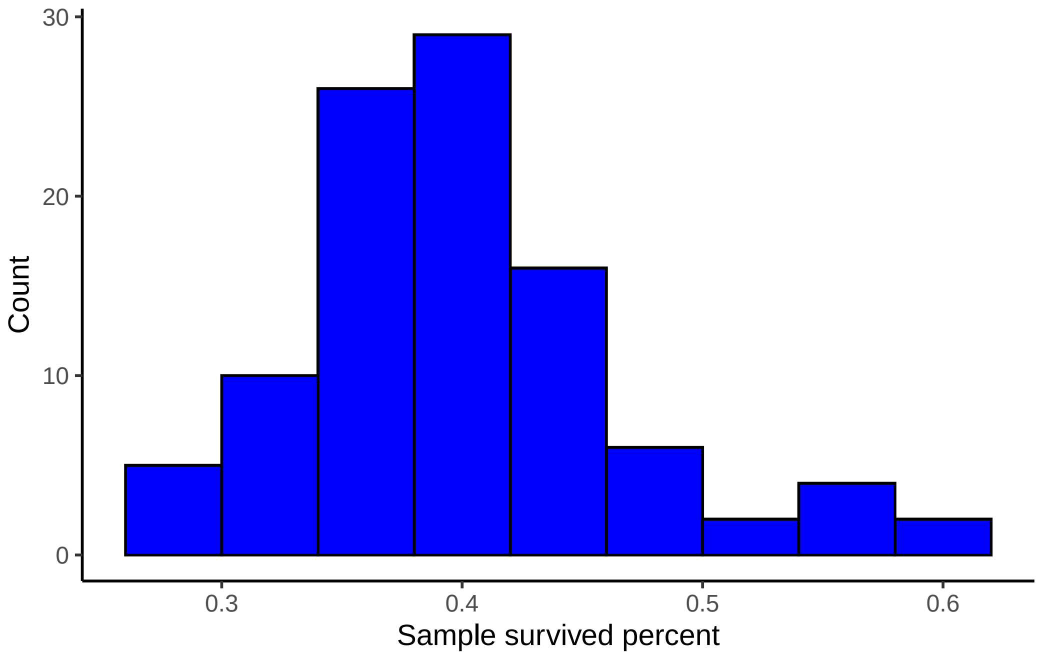
6. Make a histogram using `ggplot` of the results of `titanic_n50_s100`. What does this histogram show? Interpret this carefully.

7. Calculate the `sd()` of `titanic_n50_s100` - what is this quantity indicate? What calculation should it be equal to? Why?

```
sd(titanic.n50.s100)
```

```
[1] 0.06812466
```

This should be equal to the calculated SE.



(a) N=50, Nsample=100

Figure 1: Histogram of Titanice passenger surived percentage

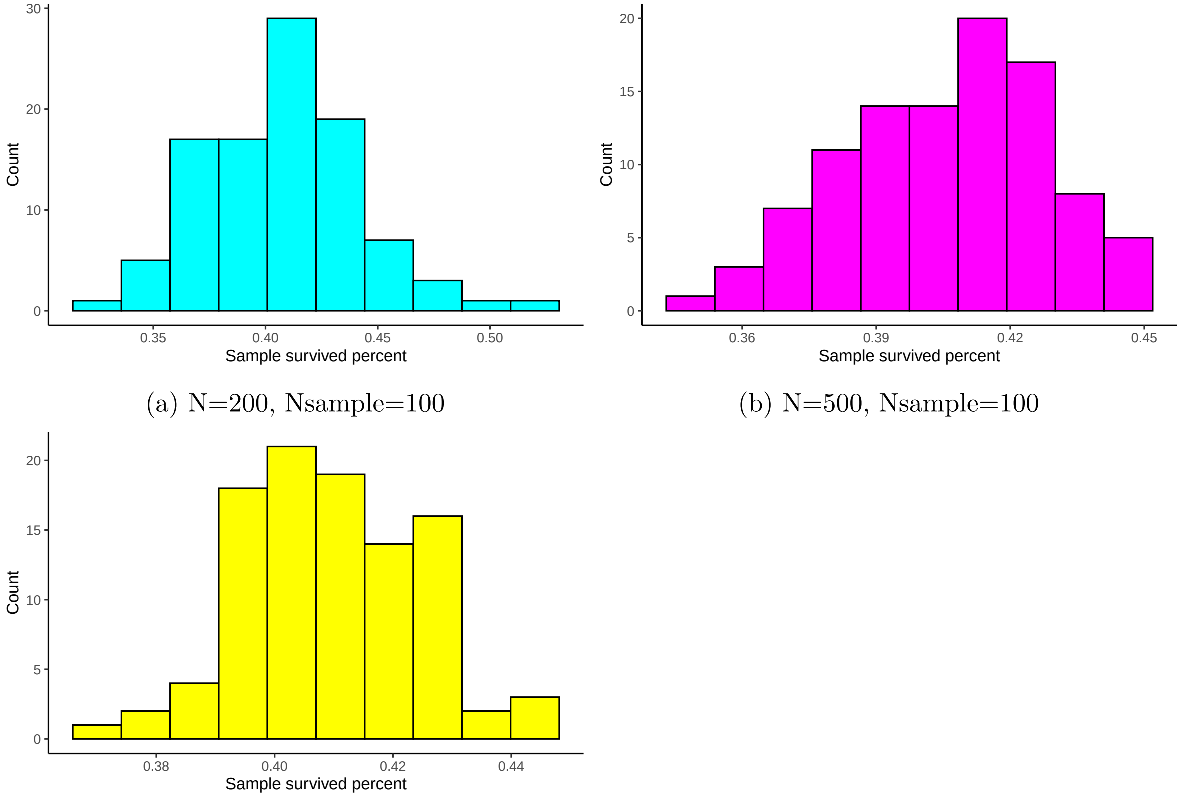
8. If you increased the number of items sampled (`n` or the first entry in `prop.multiple.samples`) what do you think will happen to your histogram? How about the `sd()` you calculated?. How about if you increased `numsamples` instead?

   The size of the standard deviation of the sampling distribution should decrase as the sample size increases (sample to sample variation should decrease). Therefore, the spread indicated by the histogram should also decrease.

   If numsamples increases, the stability of the shape of the sampling distribution should improve - it should become more obviously normal.

Comparing the distribution

9. Increase the `n` and `numsamples` separately (try values like 200 and 500 and 10000). How does the shape and distribution of the histogram change? Did it match your expectations? Why or why not?



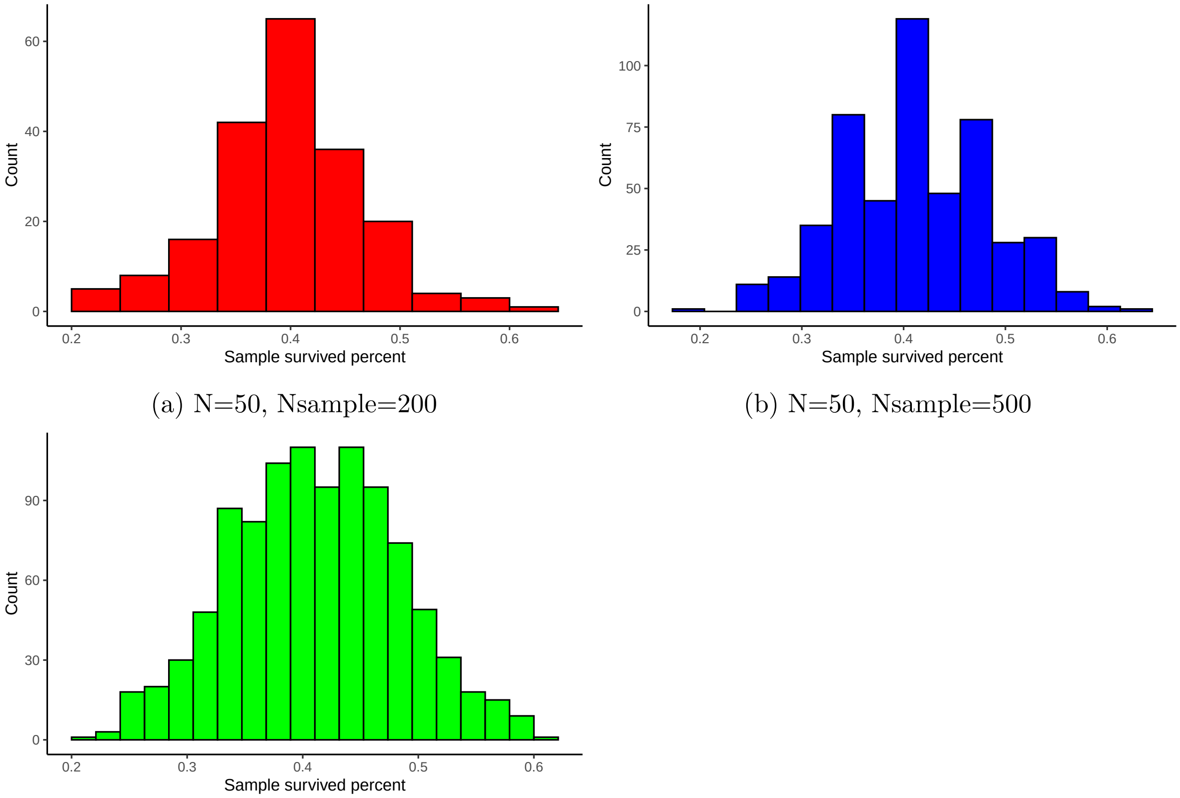
(a) N=200, Nsample=100

(b) N=500, Nsample=100

(c) N=1000, Nsample=100

Figure 2: Channging the value of sample size (N)

As expected, the distribution of the sampling distribution decreased as the sample size increased.



(a) N=50, Nsample=200

(b) N=50, Nsample=500

(c) N=50, Nsample=1000

Figure 3: Channging the value of sample size (N)

Here the difference is less obvious but the shape does appear to be more normally distributed.

Extra activity - modeling

While it is more common to use a non-linear link function (logistic regression) to model an outcome variable with a 0 or 1 outcome, in this case please try to use linear regression make a model that predicts what factors are most important in predicting survival on the Titanic.

Table 2: Logistic regression model predicting status: survived

Characteristic	**OR**	**95% CI**	**p-value**
(Intercept)	33.8	18.1, 65.2	<0.001
passengerclass			
1st	—	—	
2nd	0.28	0.18, 0.43	<0.001
3rd	0.10	0.06, 0.16	<0.001
sex			
female	—	—	
male	0.08	0.06, 0.11	<0.001
age	0.97	0.95, 0.98	<0.001

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

McFadden R-squared: 0.305; Residual Std. Error ($\sqrt{}$deviance/df): 0.972