Lecture 3.2 - Practicing Confidence Intervals II

Student

2025-11-18

Practicing Confidence Intervals

Setup

Pretend you are a real estate agent interested in renting houses in one of the cities listed in the dataset. This dataset contains the rental prices of one large firm's rental property portfolio across several major U.S. cities.

You can view the list of cities in the dataset by:

```
table(<your dataset name>$City)
```

1. Go online and search for some information about the rental market in that city and write down some expectations about the rental prices.
2. Filter for only rentals in your chosen city using the following command:

```
rental.sample %>% <your dataset name>
  filter(City=="<your chosen city>")
```

3. Take a sample of this data of size 50 using the following command:

```
rental.sample <- rental.sample %>%
  slice_sample(n=50)
```

Exploring the data

4. Make a histogram of the rental prices. Write a sentence about the distribution of rental prices in your sample. Do you think the data, based on your histogram, is suitable for a confidence interval? Why or why not?

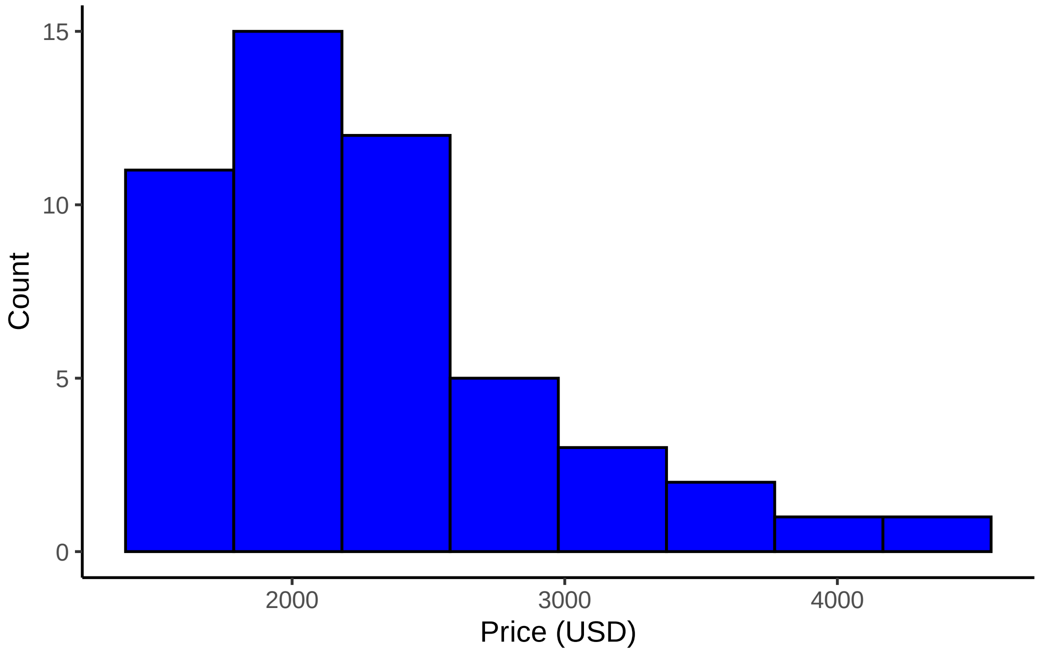

Figure 1: Distribution of Seattle rental prices for sample n=50

5. Calculate the mean and standard deviation of rental prices in your sample. Then, using a calculator, create and interpret a 95% confidence interval for the mean rental price in your sample using the formula from the textbook.

```
round(mean(seattle.apts.sample$Price), 2)
```

```
[1] 2296.38
```

```
round(sd(seattle.apts.sample$Price), 2)
```

```
[1] 635.4
```

$\bar{y} \pm t_{n-1} \cdot \frac{s}{\sqrt{n}}$

$2296.38 \pm 2.01 \cdot \frac{635.4}{\sqrt{50}}$

(2115.8, 2476.96)

Thinking about validity

6. Consider the conditions required for the confidence interval to be valid. Write a brief paragraph addressing each of the conditions and whether they are met.

   Conditions for a confidence interval:

   1. Randomization: The data was randomly sampled, but from the company's data. We would have to be careful about generalizing the data
   2. Independence: Met by the randomization condition
   3. Nearly normal - since our sample size is large enough, we should be guaranteed that the sampling distribution is normally distributed

7. Consider whether the mean accurately reflects the rental price. Write a brief paragraph describing why or why not the mean would be a useful prediction for the price you might pay for an apartment rent.

   The mean could be pulled by outliers (though this does not appear to be the case from the sample). In that case, it wouldn't represent what a 'typical' apartment would rent for.

Going from your sample to the population

8. Now find the mean of price from the subset of all rentals from the city, not your sample. Does your confidence interval cover the dataset's mean for your chosen city? How close was your sample's standard deviation to the population standard deviation? About what percent of the time would you expect the confidence interval to cover the true population mean?

```
round(mean(seattle.apts$Price), 2)
```

```
[1] 2343.81
```

Yes it is covered

9. How useful would this confidence interval be, in your opinion, for an aspiring real estate agent in your city?

It could be helpful for considering what incomes of renters to target, if there are any trends in increasing prices, etc.

10. An important statistical concept is called a *sampling frame*. What is the sampling frame of this dataset? How does it limit the utility of the findings? Why might you be careful about generalizing your result based on your calculation here?

The sampling frame of the data is all apartments from the company Equity Apartments on a particular date.

A limit on the utility of this is that Equity Apartments may not be that similar to the overall market or may offer only specific kinds of apartmnets.

If their are not that similar to the overall market, a random sample of their apartments would not represent the population very well.

Expanding on your findings

11. Create a regression model that you think would better predict rental price in your chosen city.

- Start with creating a regression model with no predictor variables using your sample data. Compare the standard error in the regression table to what you calculated by hand.

```
Call:
lm(formula = Price ~ 1, data = seattle.apts.sample)

Residuals:
   Min     1Q Median     3Q    Max 
-774.4 -443.9 -145.9  246.9 2003.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2296.38      89.86   25.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 635.4 on 49 degrees of freedom


*** The SE from the previous step was: 89.86
```

Note that the intercept is exactly the mean we found earlier and the SE of the mean is the SE we calculated earlier. You can get exactly the CI information from a regression with no predictors.

- Create a more expansive model based on your model-building skills

  Examples here will vary