Sylvia Liaw (sl33255)

February 3 Scribing

Summary

Overall, we did a quick run through of our homework, and Professor Scott said we would go over more on Wednesday. In class, we did an exercise where we all simulated "fishing," and then we learned how to do this simulation on the computer in R. The simulation involved the sampling distribution of populations. Lastly, we learned about bootstrapping, which is a way to find the sampling distribution without needing to know the population.

Went over Homework 3

1. **Problem Number 1**
    a. Max was African Elephant, Min was Lesser Tailed Schrew
        i. Use max() and min() to find
    b. Max was Asian Elephant, Min was African Elephant
        i. Use max() and min() of residuals to find
    c. Going over this one more on Wednesday
2. **Problem Number 2**
    a. (24.6462 to 129.8704)=prediction interval
        i. Used David Puelz' code from his email
            1. newdata=data.frame(body=100)
            2. predict(lm1,newdata,level=.95, interval="prediction")
    b. Chose the polynomial function with the smallest range of the prediction interval
        i. There is a tradeoff between fit and simplicity
        ii. The 3rd model was chosen
        iii. As the polynomial increases in power, the function is a closer fit
3. **Problem Number 3**
    a. We will go over this on Wednesday

Class

1. Assume there is a True State of the World where--
    a. $Y= B_0 + B_1X + error$
        i. You take a sample, you fit a model, and you find predictions for $B_0 + B_1X$
            1. These are just estimates that you hope are close to the real thing
        ii. But how sure are we of the estimates?
            1. Answer is a confidence interval
2. Class Exercise: Fishing
    a. Bucket of random pieces of paper. Everyone grabs 10 pieces of paper randomly
    b. The papers have Weight, Length, Height, and Widths of a fish
    c. Reported these numbers into Excel. Saved it as a csv.

###################################### IN R STUDIO ######################################

    d. Load the data using "import dataset"
        i. This is a random sample

e. We want an estimate of the linear relationship between weight and volume
    i. # Define volume
        1. volume=(fish$Length * fish$Width * fish$Height)
    ii. # Plot the data
        1. plot(weight~volume, data=fish)
    iii. # Get a linear relationship between volume and weight
        1. lm1=(weight~volume, data=fish)
    iv. # Add the linear model to see it on the graph
        1. abline(lm1)
    v. # Extract the coefficients
        1. coef(lm1)
f. Students all reported their coefficients, and then we aggregated all of the linear models
g. Plotted everyone's linear models (called sampling distribution) and saw that there was a lot of variation
h. Plotted a histogram of everyone's slopes
i. The amount they vary is the certainty you can have of the population
    i. Unstable estimates = less trustworthy
j. When looking at the sampling distribution, you are aggregating everyone's sample. The tighter the aggregate plot (sampling distribution), the more trustworthy your coefficients seem
    i. Standard error = standard deviation of sampling distribution

3. Now follow along in the gonefishing.R script
    a. Basically, in this script, we replicate our class fish exercise, but on the computer
    b. Line 12= equation to get the true coefficients for the whole population of fish
    c. Line 20, we changed n to 10 instead of 30 to change the number you sample
        i. Because in class we sampled 10 fish each
    d. Added in "nsamp(gonefishing.nsamp)", found the coefficients. Do this 10 times.
        i. Taking simulated, different samples of 10 fish
        ii. SHORTCUT: easier way to do this is line 29
    e. Now we can find a histogram of the 1000 samples to the volume
        i. The confidence interval is the 95% middle section of histogram
        ii. If we change our samples to 30 instead of 10, the histogram will get tighter
        iii. Tighter = smaller range = less biased
    f. Use sd() command to verify that the histogram is tighter and looks unbiased
        i. By using this command, you see how much it deviates
    g. To check where the histogram is centered, use "colMeans(montecarlo)"
        i. This function gives the column means of the intercept, volume, sigma, and rsquared of all of the samples
        ii. Look at the volume row. Gives us a very close number to the real (lmfull) answer
            1. Means our found answer was unbiased
    h. Extract the 95% confidence interval for each model parameter
        i. "confint(montecarlo, level=.95)"

ii. Gives lower and upper of the confidence interval. Be sure to look at the volume row to get these numbers
iii. Difference between coverage and confidence interval= coverage interval is everything and very generic. Confidence interval is of the sampling distribution.
iv. BUT ALL OF THIS IS USELESS
1. Too tedious. Cannot take repeated samples from population because then you have a larger sample

i. INSTEAD
i. How do things change from sample to sample?
ii. New method, called bootstrapping!!
1. Term means—to get something from nothing, or from very low circumstances

4. Bootstrapping
a. "I want to see how things change from one sample to the next when I take it from my sample"
b. Everything that was once "population," gets changed to "my sample."
c. Go to line 50 in r script
i. Plot line 56
ii. Add in the line as well
d. Bootstrapping is taking a sample of your sample, instead
i. Wouldn't it be the same each time you take a sample of your sample? Not if you sample with replacement
1. Each time you take a fish, you throw it back in before you take another. This means there is possibility for repetition and omission.
e. Called bootstrapped resamples, and fitting a linear model to it
f. Do it 1000 times. Store it as myboot
g. View the histograms. There is no population here. No need for access to full population
h. This is not the "true" sampling distribution
i. This is an estimate of the sampling distribution using bootstrapping
ii. We hope that this one is about as spread out as the true one
iii. Look at the estimate of the standard error of our bootstrap sample and standard error of population to compare the distributions