Due 11/20 at 11:59 PM ET

Assignment 5 (65 pts)

1. Recall: We looked at how the width of the confidence interval would change if you had a different sample size (Week 4, Lecture 2, slide 30). Using the same example, explore how the confidence interval would change if the standard deviation were different.

Recall, sample mean ()=261.9, sample standard deviation(s)=95.6, and sample size was 308. The original 95% confidence interval is 251.2 to 272.7 *(Hint: the t critical value is 1.97)*

* 1. Suppose the standard deviation was 60. Recalculate the 95% confidence interval and interpret it (5 pts)

Our sample of 308 patients had a mean () of 261.9, and a SD (s) 60.

* + 95% CI: 251.2, 272.7 (width: 21.5)
  + T critical value=1.97
  + 95% CI: =261.9±1.97\*( ) =261.9±6.735
  + 95% CI: 255.165, 268.635 (width 13.47)

**Interpretation**

Compared to the width of 21.5 with a s=95.6, the width has decreased to 13.47 with s=60. As the standard deviation decreases the width of 95% confidence interval decreases.

* 1. Suppose instead that the standard deviation was 150. Recalculate the 95% confidence interval and interpret it (5 pts)

Our sample of 308 patients had a mean () of 261.9, and a SD (s) 150

* + 95% CI: 251.2, 272.7 (width: 21.5)
  + T critical value=1.97
  + 95% CI: =261.9±1.97\*( ) =261.9±16.837
  + 95% CI: 245.063, 278.737 (width 33.674)

**Interpretation**

Compared to the width of 21.5 with a s=95.6, the width has increased to 33.674 with s=150. As the standard deviation increases the width of 95% confidence interval increases.

Standard deviation (s) is directly proportional to 95% confidence interval because as former (s) increases the later (CI) increases and as former (s) decreases later (CI) decreases.

* 1. Recall: We did a 1 sample t-test to compare the mean platelet count for PBC patients to the general population (275). We rejected the null hypothesis and concluded the mean for PBC patients was different from 275. Based on the results from part b, would that decision change if the standard deviation had been 150 instead of 95.6? Why or why not? (5 pts)

95% CI for Standard deviation 150: 245.063, 278.737

H0: μ=275

H1: μ≠275

The range of plausible values for the population mean includes 275. Based on the 95% confidence interval, 275 is a mean that we would consider plausible for PBC patients, given what we saw in our data

1. Use the *pbc\_Mayo Clinic 312 pats.csv* file from last week to answer the following questions. Recall the 1 and 2 sample t-tests about serum albumin levels (1b and 1e on Assignment 3).
   1. Using SAS, rerun the 1 sample t-test from Assignment 3 (#1b) to get the 95% confidence interval for the mean serum albumin level in the population of PBC patients and interpret. Using the confidence interval, describe the decision and conclusion of the 1 sample t-test, comparing serum albumin for PBC patients to 4.4 (5 pts)

Null Hypothesis: Population mean serum albumin level for PBS patients is equal to 4.4

Alternate Hypothesis: Population mean serum albumin level for PBS patients is not equal to 4.4

H0: μ=4.4

H1: μ≠4.4

Table

Description automatically generated Text

Description automatically generated with medium confidence

P-value=<0.0001

95% CI= 3.4732, 3.5668

**Decision/Conclusion for P-value**

Since this result is very unlikely if the null hypothesis were true (P-value<0.05), we reject the null hypothesis. We conclude that the mean albumin count for the population of PBS patients is statistically significantly lower than 4.4

**Decision/Conclusion for CI**

We are 95% confident that the population mean albumin value for PBC patients is between 3.4732 - 3.5668. The range of plausible valuesfor the population mean does not include 4.4. Based on the 95% confidence interval, 4.4 is not a mean that we would consider plausible for PBC patients, given what we saw in our data.

* 1. Using SAS, rerun the 2-sample t-test from Assignment 3 (#1e) to get the 95% confidence interval for the difference of the means of serum albumin in the treated and untreated groups and interpret. Using the confidence interval, describe the decision and conclusion of the 2 sample t-test comparing mean albumin levels for the treated and untreated PBC patients. (5 pts)

**Null Hypothesis:** The difference in the mean albumin count of patients who received D-penicillamine (treatment group) and who did not receive D-penicillamine (placebo group) is equal to zero

**Alternate Hypothesis:** The difference in the mean albumin count of patients who received D-penicillamine (treatment group) and who did not receive D-penicillamine (placebo group) is not equal to zero

H0: μ1= μ2 (μ1-μ2 =0)

H1: μ1 ≠μ2 (μ1-μ2 ≠ 0)

Graphical user interface

Description automatically generated with medium confidence

Table

Description automatically generated

Two Sample T-Test:

* P-Value for equality of Variances= 0.1602
* Here our P-Value is >0.05
* If P-Value is greater than 0.05 we can assume that our Variance is equal, and we will use the Pooled method
* Therefore P-Value=0.8730

**95% Confidence Interval**

* trt1=3.5163, 95% CI: 3.4466-3.5859
* trt2=3.5238, 95% CI: 3.4808-3.5868
* trt1-trt2=-0.00757, 95% CI: (-0.1013) -(-0.00861)

**Decision/Conclusion for P-value**

Since the result is not that unlikely if the null hypothesis were true (P-Value>0.05), we do not reject the null hypothesis. We don’t have the sufficient evidence to conclude that the Mean Albumin count for PBC patients who received D-penicillamine (treatment group) and who did not receive D-penicillamine (placebo group) are different.

**Decision/Conclusion for CI**

We are 95% confident that the difference of the means of serum albumin in the treated and untreated groups is somewhere between (-0.1013) - (-0.00861). The range of plausible values for the differences in two groups does not include zero (0), the difference in means is -0.00757. Based on the 95% confidence interval, is μ1-μ2 =0 is not a mean that we would consider plausible for PBC patients, given what we saw in our data.

1. Choose the correct answer: As Power increases, the probability of a type II error: (1 pts)
   1. Increases
   2. **Decreases**
   3. There is no relationship between power and the Type II error

As power increases the probability of type II error decreases.

1. In a couple of sentences, discuss why it is important to make sure your study has an adequate sample size (4 pts)

* To get more power for detecting differences and to minimize type II errors we need a larger sample size. If you want more power to make correct decisions, and to detect the differences whether they exist or not we need more patients (larger sample) to do so.
* We need larger sample size to overcome larger standard deviation in order to maintain the required power
* The smaller the difference you are trying to detect, the more difficult it is to see whether that difference exists in the data. If smaller the effect size to detect we need the larger sample size.
* Larger is the sample size better are the results of the study

1. Use the *pbc\_Mayo Clinic 312 pats.csv* to answer the following questions. We are interested in the relationship between age and serum albumin, as liver function decreases with age.
   1. Plot a scatterplot of for age and albumin. Describe the relationship (5 pts)

Chart, scatter chart

Description automatically generatedGraphical user interface, text

Description automatically generated

The scatterplot appears to have negative linear relationship

between age and albumin

* 1. Report the correlation (r) between age and serum albumin and interpret. Comment on whether the correlation is significantly different from 0 (5 pts)

Table

Description automatically generatedGraphical user interface, text

Description automatically generated

r= -0.19526

P= 0.0005

Since our P-value is 0.0005 which is different from 0, there is some linear relationship. There is statistically significant week or small negative linear relationship between age and serum albumin (r=-0.19526).

* 1. Run a simple linear regression to predict serum albumin from age. Report and interpret the slope and 95% confidence interval. Comment on whether the slope differs significantly from 0 (5 pts)

Age= predictor variable/independent variable

Graphical user interface, text, application

Description automatically generatedAlbumin=outcome/response/dependent variable

Table

Description automatically generated

B1 (slope)= -0.00775

95%CI= -0.01210, -0.00340

P-value= 0.0005

Comment

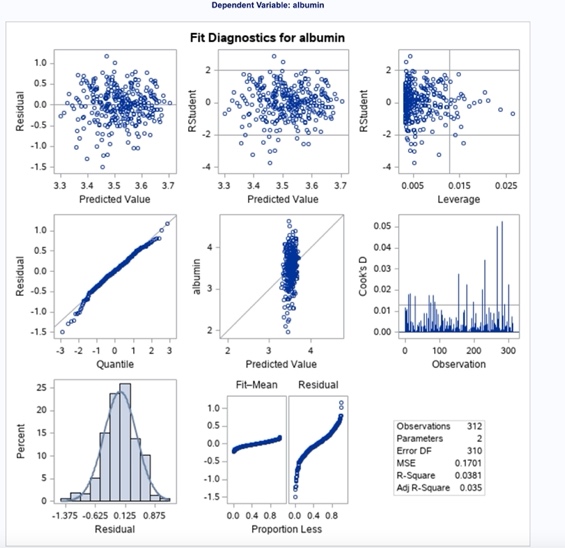
* P on B1 is <0.05, we do not reject null hypothesis B1≠0.If B1≠0, then for every 1-unit increase in age, on average albumin changes by -0.00775
* Since our slope B1 is not equal to zero, there is linear relationship between age and albumin. The slope is significantly different from 0 (p<0.0005), for every 1-unit increase in age, on average albumin changes by -0.00775. We have negative relationship between age and albumin, as age increases, albumin decreases by 0.00775. We are 95% confident that the population slope is between -0.01210 and -0.00340.
  1. Identify and interpret the coefficient of determination (R2). Does age explain changes in serum albumin well? (5pts)

R2=0.0381

Goodness of fit p-value=0.0005

There is significant linear relationship. 3.81% of variability in albumin is explained by age. Only 3.81% of difference in albumin are due to age.

* 1. Look at the plots of residuals. Using the plots shown in the lecture, discuss whether we meet the assumptions for linear regression in 2-3 sentences. (5 pts)



1st residual plot:

* Randomly dispersed around zero
* No obvious pattern

4th and 7th residual plots:

* Dots along diagonal
* No serious departure from normality
  1. Compare the multiple regression output in the ‘Assignment 5 Multiple Regression Output’ (full output in ‘Assignment 5 model results.html’ file), to predict serum albumin from age, presence of hepatomegaly or enlarged liver (*hepato),* bilirubin (*bili*), and standardized blood clotting time (*protime*).Which model would you select and why. Interpret all of the slopes and 95% confidence intervals for each variable and the R2 for the final model (10 pts) *(Review the pbc data dictionary for interpretation of values of each variable; See full SAS output and summary files)*

**Model 1: Serum albumin from age:**

**P=** 0.0005

**B1**= -0.008

**95% CI**= -0.012 to -0.003

**R2** = 0.381

**Goodness of fit p-value**=0.0005

**Interpretation**

* If our slope (B1)≠0, then for every 1-unit increase in age, on average albumin changes by -0.008.
* Since our slope B1 is not equal to zero, there is linear relationship between age and albumin. The slope is significantly different from 0 (p<0.0005), for every 1-unit increase in age, on average albumin changes by -0.008. We have negative relationship between age and albumin, as age increases, albumin decreases by 0.008. We are 95% confident that the population slope is between -0.01210 and -0.00340.
* Age: The slope for age (B1) is significantly different from 0. For every one-year increase in age, albumin decreases by 0.008 accounting for hepatomegaly, bilirubin and protime.
* 3.81% of variability in albumin is explained by age. Only 3.81% of difference in albumin are due to age.

**Model 2: Serum albumin from age and hepatomegaly:**

|  |  |  |
| --- | --- | --- |
|  | Age | Hepatomegaly |
| P | 0.0013 | <0.0001 |
| B1 | -0.007 | -0.24 |
| 95% CI | -0.01 to -0.003 | -0.33 to -0.15 |
| Goodness of fit P-value | 0.0013 | <0.0001 |
| VIF | 1.01 | 1.01 |

**Interpretation for age**

Since our slope B1 is not equal to zero, there is linear relationship between age and albumin. The slope is significantly different from 0 (p<0.0005), for every 1-unit increase in age, on average albumin changes by -0.008. We have negative relationship between age and albumin, as age increases, albumin decreases by 0.008. We are 95% confident that the population slope is between -0.01210 and -0.00340.

**R2**= 0.1138

* This model does a little bit better than at predicting albumin than the previous model (**R2**= 0.1138).
* All VIF<5, no serious multicollinearity
* This model better explains FEV, and all slopes are significantly different from ‘0’.

**Model 3: Serum albumin from age, hepatomegaly and bilirubin**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Age | Hepatomegaly | Bilirubin |
| P | 0.0011 | 0.0002 | <0.0001 |
| B1 | -0.07 | -0.17 | -0.02 |
| 95% CI | -0.01 to -0.003 | -0.26 to -0.08 | -0.03 to -0.01 |
| VIF | 1.01 | 1.11 | 1.10 |
| Goodness of fit P value | 0.0011 | 0.0002 | <0.0001 |

**R2**= 0.1758

**Model 4: Serum albumin from age, hepatomegaly, bilirubin and protime**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Age | Hepatomegaly | Bilirubin | Pro-time |
| P | 0.0033 | 0.0003 | <0.0001 | 0.1604 |
| B1 | -0.006 | -0.17 | -0.02 | -0.03 |
| 95% CI | -0.01 to -0.002 | -0.26 to -0.08 | -0.03 to -0.01 | -0.08 to 0.01 |
| VIF | 1.04 | 1.11 | 1.23 | 1.20 |
| Goodness of fit P-value | 0.0033 | 0.0003 | <0.0001 | 0.1604 |

**R2**= 0.1784