Use the Height tab. This dataset has 4 variables and *n* = 56 observations:

*female*: has values ‘1’ for females and ‘0’ for males

*sleep*: the ‘typical’ amount of sleep for the person (in hours)

*shoe*: the American shoe size of the person

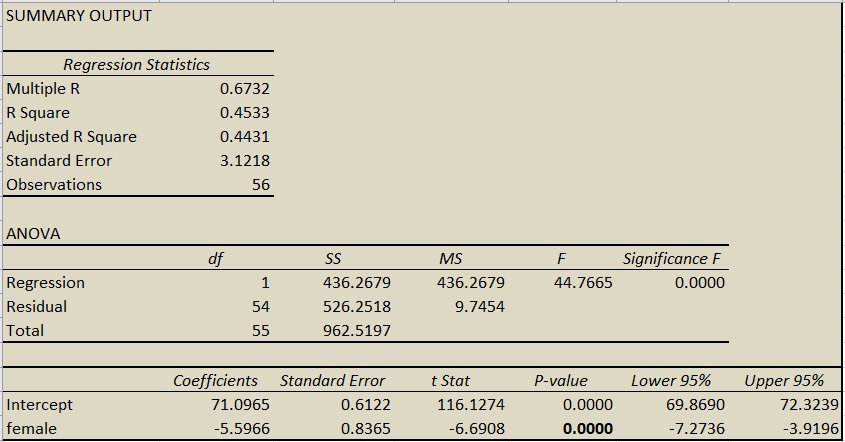
*height*: the height of the person (in inches) Use this dataset to answer the following questions:

1. Run a simple regression of *Y* = *height* vs. *X* = *female*. Perform a hypothesis test to see if the variables are associated. What is the interpretation of the slope estimate, *b*1?

Answer a)

Null hypothesis: β1 = 0 corresponds to the mean value of Y is the same for both groups.

Alternative hypothesis: β1 not equals 0, the mean value of Y is not the same for both groups.



The interpretation of slope coefficient of β1 is -5.6, implies that each increase of 1 person in female, then value of height is estimated to decrease by 5.6 cm.

Thus, the estimated height of females is 65.5 cm, while the estimated height of males is 71.1cm.

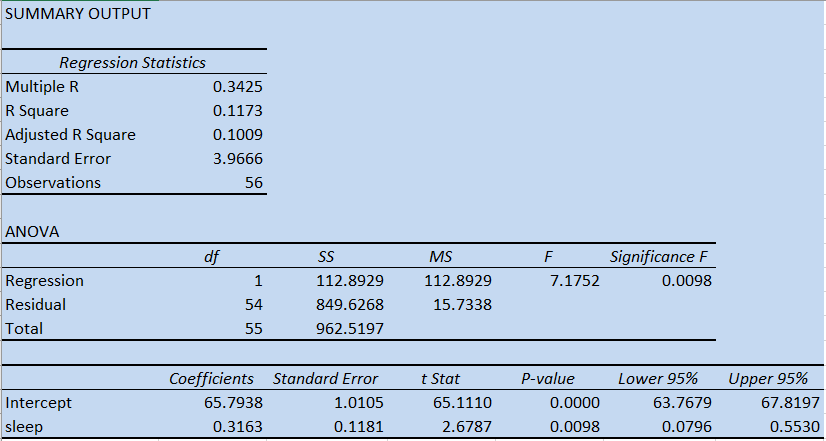
Moreover, the p-value of the female is close to zero which is less than the level of significance (0.05). We can easily reject the null hypothesis based on the p-value and support an alternative hypothesis. Therefore, the mean value of Y is not the same for both groups.

1. Run a simple regression of *Y* = *height* vs. *X* = sleep. Perform a hypothesis test to see if the variables are associated.

Answer b)

Null hypothesis: β1 = 0 corresponds to the mean value of Y is not associated with sleep.

Alternative hypothesis: β1 not equals 0, the mean value of Y is associated with sleep.



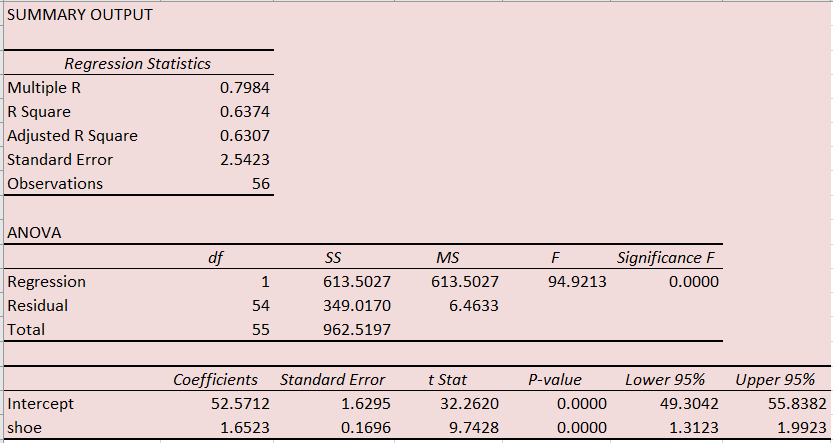
The p-value tells whether the null hypothesis is true or not.So, the p-value of the sleep variable is 0.0098 which is less than the level of significance. In this case, we can reject the null hypothesis and support an alternative hypothesis. Thus, the result is that there is a statistically significant relationship between sleep, and height.

1. Run a simple regression of *Y* = *height* vs. *X* = *shoe*. Perform a hypothesis test to see if the variables are associated.

Answer c)

Null hypothesis: β1 = 0 corresponds to the mean value of Y is not associated with shoes.

Alternative hypothesis: β1 does not equals 0, the mean value of Y is associated with shoe.



The p-value tells whether the null hypothesis is true or not. So, the p-value of the sleep variable is close to zero (=0.0000) which is less than the level of significance. In this case, we can reject the null hypothesis and support an alternative hypothesis. Thus, the result is that there is a statistically significant relationship between the shoe, and the height.

1. Run a multiple regression of *Y* = *height* vs. *X*1 = *female* and *X*2 = *shoe*. Perform a single hypothesis test to see if either of the *X*-variables are associated with height.

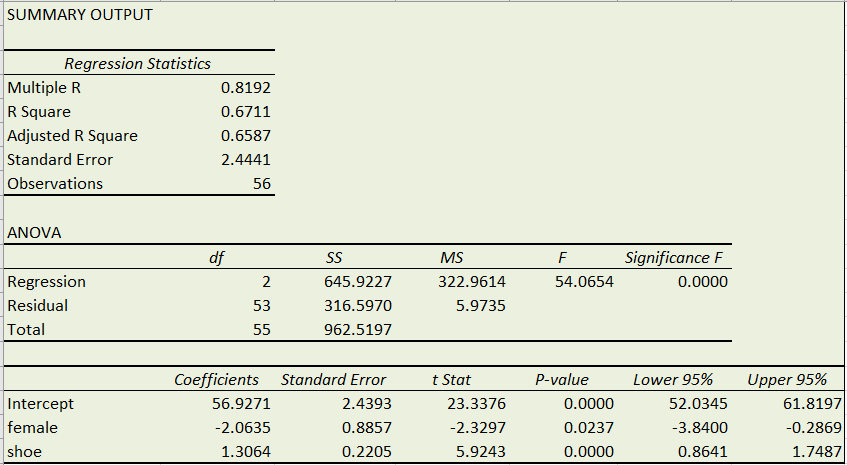
Answer d)

Null hypothesis: β1 = 0 corresponds to the mean value of Y is not associated with females.

Alternative hypothesis: β1 does not equals 0, the mean value of Y is associated with females.

Null hypothesis: β2 = 0 corresponds to the mean value of Y is not associated with shoe.

Alternative hypothesis: β2 does not equals 0, the mean value of Y is associated with shoe.



The p-value of the female and shoe is close to zero which is less than the level of significance (0.05). Thus, both the variables are statistically significant and we can reject the null hypothesis and in favor of the alternative hypothesis. We can easily conclude that both the variables are a statistically significant relationship between the shoe & the height and female and height. On the other hand, the R-square of the model is more than 80% which indicates that the model is a good fit.

1. Run a multiple regression of *Y* = *height* vs. *X*1 = *female, X*2 = *shoe,* and *X.*= *sleep*. Perform a single hypothesis test to see if any of the *X*-variables are associated to *height*.

Answer e)

Null hypothesis: β1 = 0 corresponds to the mean value of Y is not associated with females.

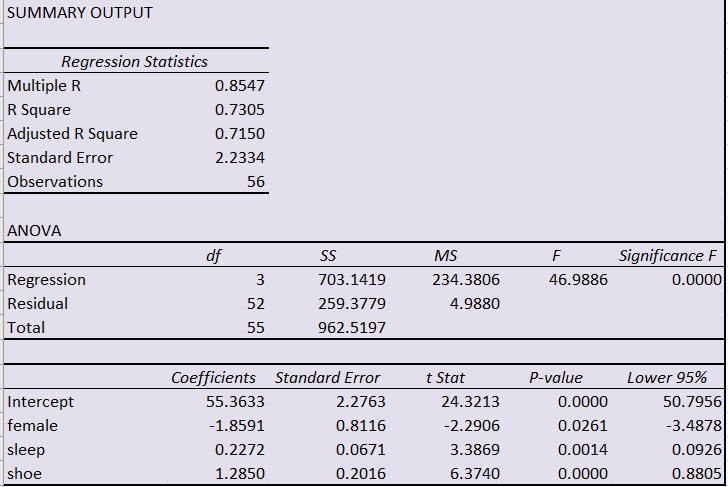
Alternative hypothesis: β1 does not equals 0, the mean value of Y is associated with females.

Null hypothesis: β2 = 0 corresponds to the mean value of Y is not associated with sleep.

Alternative hypothesis: β2 does not equals 0, the mean value of Y is associated with sleep.

Null hypothesis: β3 = 0 corresponds to the mean value of Y is not associated with shoe.

Alternative hypothesis: β3 does not equals 0, the mean value of Y is associated with shoe.



The p-value of all variables close to zero which is less than the alpha level (=0.05). Thus, all the null hypotheses are rejected. Thus, female, sleep, and shoe variables are statistically significant and we can reject the null hypothesis and in favor of the alternative hypothesis. We can easily conclude that all the variables are a statistically significant relationship between female and height, shoe & height , sleep and height. On the other hand, the R-square of the model is around 85% which indicates that the model is a good fit.

1. Using you multiple regression model in part (e), interpret the coefficient, *b*1, for the variable *female*.

According to the regression model of output e, the interpretation slope coefficient of b1 = -1.86, implies that each increase of 1 person in female, then value of height is estimated to decrease by 1.86 cm.

1. Using you multiple regression model in part (e), make a prediction of your height based on the values of the predictors for you (sex, shoe size, and typical amount of sleep). Be sure to clearly label what your *x*-values are.

Answer g) Here is the regression below

=

X-values:

Female = 1

sleep = 5 hours

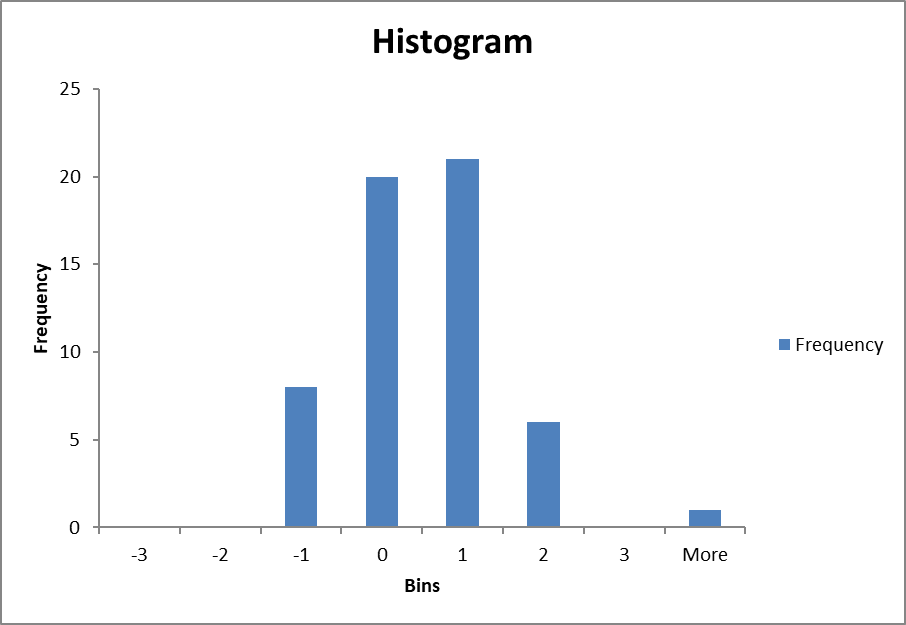
Shoe = 8

So,

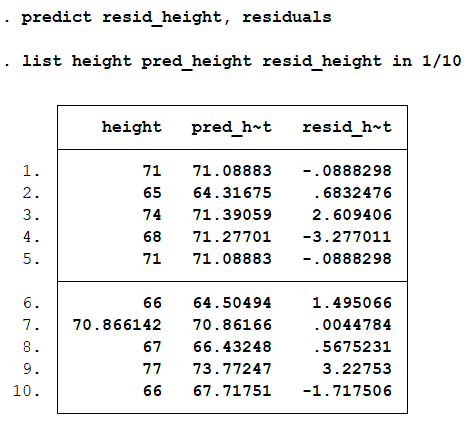
=

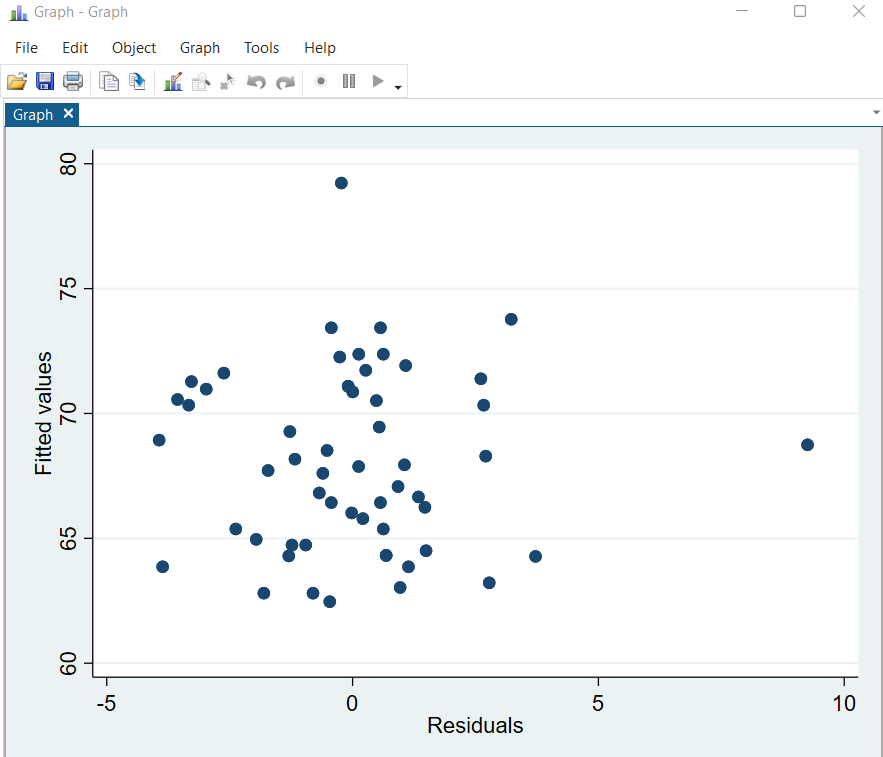
Thus, = 64.916 cm.

1. Create two plots for the multiple regression model from part (e): 1) the histogram of the residuals and



1. the residuals-versus-fitted scatterplot (remember: you can create the residuals variable using the “predict residuals, resid” command in Stata). Use these plots to comment on the validity of the assumptions for this model (be sure to include these plots).





Stata Codes:

**regress height female sleep shoe**

**predict pred\_height**

**list height pred\_height in 1/10**

**predict resid\_height, residuals**

**list height pred\_height resid\_height in 1/10**

**scatter pred\_height resid\_height**

* 1. Compare your results in parts (a) through (e) [**using a table to organize your answer is fine**]. Make sure to compare R2 values and note any major changes in significance and/or signs of slopes, especially that related to the variable *female*. Which model would be your choice as a ‘best predictive model’?

Comparison from (a) to (e)

| **Model** | **Coefficients (DV = Height)** | **R-Square** | **Significant, (p-value < 0.05)** |
| --- | --- | --- | --- |
| **Model 1 (Part a)** | Female = -5.5966 | 0.6732 | Significant |
| **Model 2 (Part b)** | Sleep = 0.3163 | 0.3425 | Significant |
| **Model 3 (Part c)** | Shoe = 1.6523 | 0.7984 | Significant |
| **Model 4 (Part d)** | female = -2.0635  shoe = 1.3064 | 0.8192 | Significant  Significant |
| **Model 5 (Part e)** | female = -1.8591  sleep = 0.2272  shoe = 1.2850 | 0.8547 | Significant  Significant  Significant |

The above table shows that the results for part a to part e. As we see that there is no change in sign of the slope of the female variable but magnitude change. If we talk about the variable significance, look at the last column, “**Significant, (p-value < 0.05)”**, for all the models, all variables are significant because variables p-value is less than the level of significance.

On the other hand, there are many R-squares, but according to the rule, the higher the R-squared, the better the model fits your data. Thus, the model 5 is the better or best predictive model.

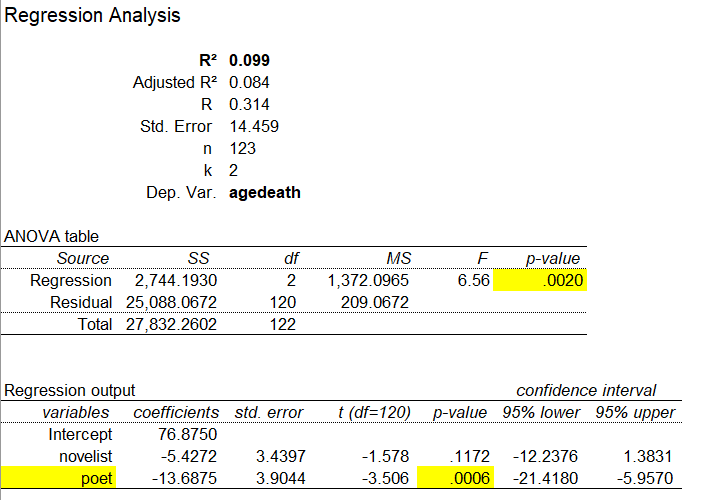
The R-square of model 5 is 85%, which implies that the model explains all the variability of the response data around its mean. In short, 85% of the variation in the height (cm) variable can be explained by the variation in the sex, sleep, and shoe variables.

1. A study was designed to determine whether different types of writers die at different ages on average. Three categories of writers were examined: novelists, poets and nonfiction writers.

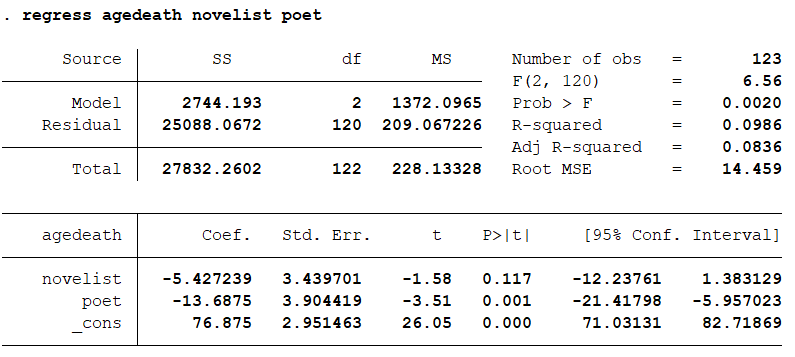
The age of death for these North American writers are found in the dataset ‘writers.csv.’ For this dataset, run a binary multiple regression using the variables *novelist* and *poet* as predictors of age of death.

1. Write out the regression model statement (in terms of *Y*, *β*’s, and *X*’s). Be sure to label what each of the variables represents.

In the below regression model, the dependent variable is age of death and explanatory variables are novelist and poet which are in dichotomous nature.



OR STATA

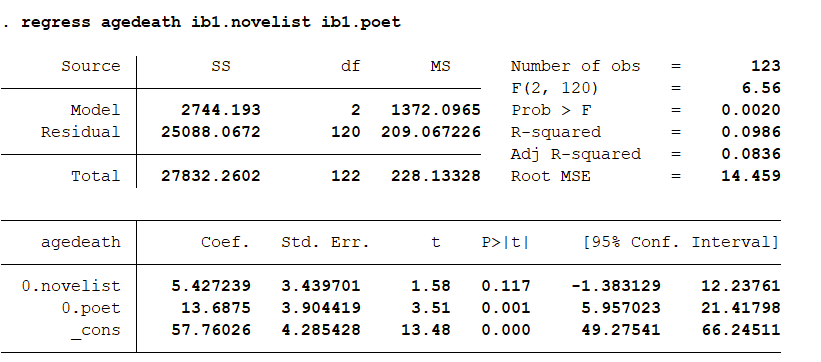


Regression equation of the model is:

Thus, the estimated age of death is 58 approximately years when novelist and poet belong to 1.

1. What is the reference group? What is the estimated difference in age at death between poets and non-fiction writers? How about between poets and novelists?

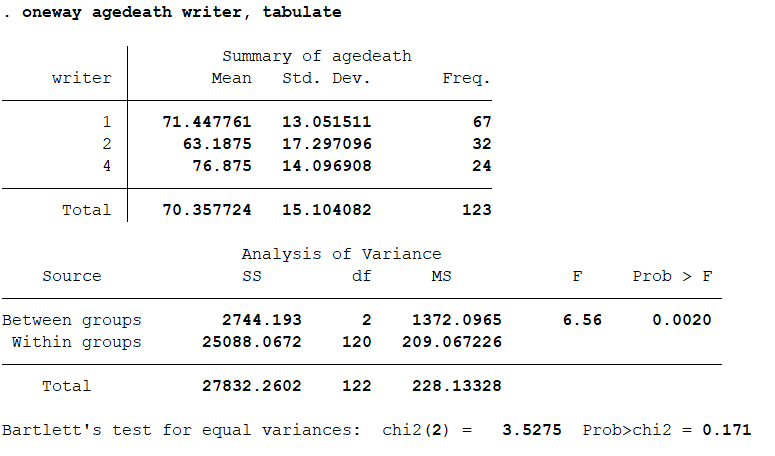
The value of the categorical variable that is not represented explicitly by a dummy variable is called the reference group.



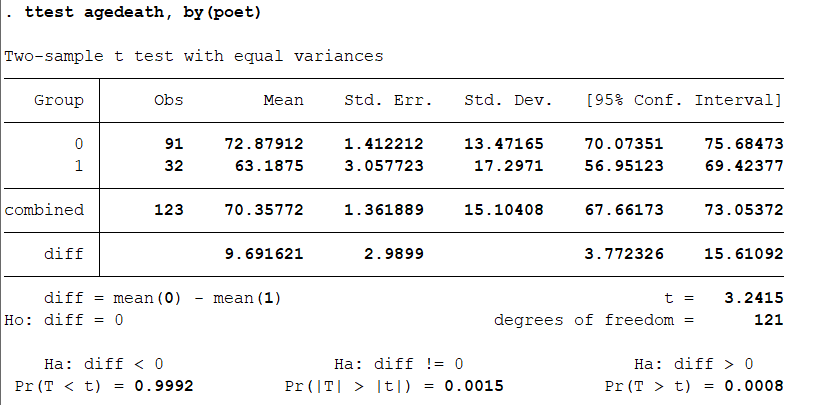
In the reference group, the regression model is the same because both variables take value 1 & 0. If both takes 1, then the slope coefficient sign is negative, however, if we run the regression based on the zero then the sign changes because the sign does not matter.

1. Perform the appropriate *F*-test to determine whether there is a difference in average age of death for the 3 types of writers.

The F-statistic is 6.56 and the corresponding p-value is 0.0020. Since the p-value is less than alpha = 0.05, we can reject the null hypothesis that the mean change in age of death for each group is equal. In other words, there is a statistically significant difference in the mean change in age of death between at least two of the writer groups.



1. Poet William Butler Yeats once wrote, “She is the Gaelic muse, for she gives inspiration to those she persecutes. The Gaelic poets die young, for she is restless, and will not let them remain long on earth.” Using the “test” command in Stata, perform an appropriate test to determine whether the average age of death for poets is different than the other 2 groups combined (you can weigh the other two groups equally or based on their sample sizes).



We can see that the group means are statistically significantly different as the p-value in the Pr(|T| > |t|) row (under Ha: diff != 0) is less than 0.05 (i.e., based on a 2-tailed significance level). Looking at the Mean column, you can see that those people who belong to group 1 of poets had lower age of death as compared to those who belong to 0.

1. For parts (a) through (d )below, **set-up** a test of hypothesis using a 0.05 level of significance. Your tests of hypothesis set-up should include:

1) a statement of the type of procedure (one- sample t-test for means, chi-squared test for association, etc…),

2) a statement of your null and alternative hypothesis,

3) whether you are performing a one or two-sided test, and

4) the general formula for the test statistic (**no need to calculate it or plug in the numbers**). If you are having difficulty deciding the type of procedure for a problem, it’s recommended to consult the *Roadmap to Inference* document on the course website.

1. The nutrition label on Kellogg’s Rice Krispies cereal claims that their product contains 3 grams of sugar per serving. To test this claim a consumer research organization conducted tests on a random sample of 16 boxes of Kellogg’s Rice Krispies cereal from grocery stores around the US. The tests showed an average of 3.25 grams of sugar with a standard deviation of 0.4 grams for the 16 servings tested. Does this consumer research organization study support the claim by Kellogg’s?

Answer a)

1. Z-test
2. Hypothesis:

Null H0:

Alternative H1:

Z-test Calculation:

Z score = (x - μ)/σ

= (3.25 - 3)/ 0.4

= 0.625

Z Score at 0.05 = 1.645

We can see that 0.625 < 1.645, thus our test statistic is in the rejection region. Therefore, we fail to accept the null hypothesis. Also, there is not enough evidence to reject H₀ at the significance level 0.05, because your p-value is greater than 0.05 < 0.5189. This consumer research organization study does not support the claim by Kellogg’s.

1. Carbon monoxide emissions are measured every 2 years for automobiles in Massachusetts during mandatory vehicle inspections. Carbon Monoxide levels (CO grams per mile) were collected for 15 Fords, 20 Hondas, and 30 Toyotas. The mean and standard deviation of the Ford CO levels were 14.2 and 4.2, for the Hondas were 11.9 and 4.1, and for the Toyotas were 12.3 and 4.3. Is this sufficient evidence of a difference in CO emissions between these three types of autos?

b) 1. ANOVA

2. The following null and alternative hypotheses need to be tested:

Null (H1):

Alternative (H2):

Ford (n1) = 15, Hondas (n2) = 20, and Toyotas (n3) = 30.

Mean (Ford) = 14.2 and S.D (Ford) = 4.2

Mean (Hondas) = 11.9 and S.D (Hondas) = 4.1

Mean (Toyotas ) = 12.3 and S.D (Toyotas) = 4.3

Overall Mean = (4.2+4.1+4.3)/3 = 4.2

ANOVA test:

We calculate that SSR (sum of squares) = 15(4.2-4.2)^2+20(4.1-4.2)^2+30(4.3-4.2)^2 = 0.5

So, SSR = 0.5

1. The Civil War Battle at Antietam was a one-day battle on September 17, 1862

near Sharpsburg, Maryland. Antietam was the bloodiest single day of battle in American history. It resulted in more than 22,000 American casualties, more than three times as many casualties than on June 6, 1944 – D-Day, the so-called “longest day” of World War II. Battle casualties are classified into three categories: 1) dead, 2) wounded, and 3) missing or captured.

The approximate number of casualties at Antietam on September 17, 1862 is given below. Is there evidence that casualty type is associated with the Army?

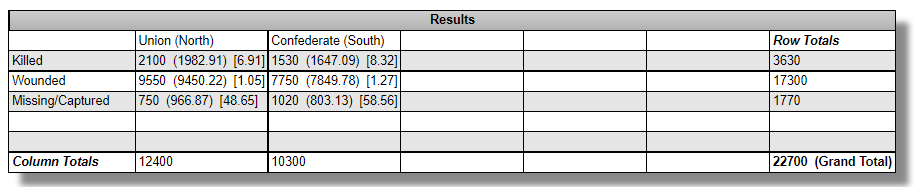
|  | Union (North) | Confederate (South) | Total |
| --- | --- | --- | --- |
| Killed | 2100 | 1530 | 3630 |
| Wounded | 9550 | 7750 | 17300 |
| Missing/captured | 750 | 1020 | 1770 |
| Total | 12400 | 10300 | 22700 |

1. Test: Chi-Square Test
2. Hypothesis:

Null H0: Casualty type is not associated with Army

Alternative H1: Casualty type is associated with Army

1. Test Statistics



Thus, the chi-square statistic is 124.7685. The p-value is < 0.00001. The result is significant at p < .05. Based on the

evidence, we can conclude that casualty type is associated with the Army.

1. Government agents in Massachusetts wanted to see if older cars have worse carbon monoxide emissions than newer cars. During mandatory vehicle inspections, carbon monoxide levels (CO grams per mile) were measured for 50 cars along with the ages of the cars (in years). The mean and standard deviation of CO levels were 13.2 and 4.2, and 4.5 and 2.3 for the age of the cars in years, with a correlation of 0.32 between CO and age. Is there evidence that older cars have increased CO emissions?

d)

Sample Mean 1 (X bar) = 13.2

Std Dev 1 = 4.2

n1 = 50

Sample Mean 2 (X bar) = 4.5

Std Dev 2 = 2.3

n1 = 50

The following null and alternative hypotheses need to be tested:

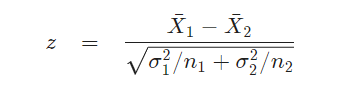
Null (H1):

Alternative (H2):

This corresponds to a two-tailed test, and a z-test for two means, with known population standard deviations will be used.

Rejection Region: Z = 1.96

Test Statistics:



Putting the values in the above formula and we get, z = 12.814

Since it is observed that |z| = 12.847 > z = 1.96(critical value), it is then concluded that the null hypothesis is rejected. Thus, the p-value is p=0.00, and since p = 0.00 < 0.05, it is concluded that the null hypothesis is rejected. Therefore, there is an evidence that older cars have increased CO emissions.