Use data set (attachment above) to calculate the following:

1. Calculate appropriate descriptive statistics (Percentage, Medians, Means, Variance, Standard Deviations) for the following variables. Calculate 95% confidence intervals for the interval-ratio variable.

first, we can calculate descriptive statistics only for quantitative data. So only AGE, Education, ARR, SELF\_CON and PCL are qualified. The results from excel are:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AGE | |  | | Education | |  | | ARR |  |
| Mean | | 33.80952381 | | Mean | | 11.19047619 | | Mean | 5.333333333 |
| Standard Error | | 2.548219778 | | Standard Error | | 0.466083212 | | Standard Error | 0.673771663 |
| Median | | 31 | | Median | | 11 | | Median | 6 |
| Mode | | #N/A | | Mode | | 12 | | Mode | 7 |
| Standard Deviation | | 11.67741002 | | Standard Deviation | | 2.135861597 | | Standard Deviation | 3.087609647 |
| Sample Variance | | 136.3619048 | | Sample Variance | | 4.561904762 | | Sample Variance | 9.533333333 |
| Kurtosis | | 0.054973283 | | Kurtosis | | -0.1749607 | | Kurtosis | -0.833640044 |
| Skewness | | 0.773601358 | | Skewness | | 0.304679229 | | Skewness | 0.329596892 |
| Range | | 44 | | Range | | 8 | | Range | 10 |
| Minimum | | 18 | | Minimum | | 8 | | Minimum | 1 |
| Maximum | | 62 | | Maximum | | 16 | | Maximum | 11 |
| Sum | | 710 | | Sum | | 235 | | Sum | 112 |
| Count | | 21 | | Count | | 21 | | Count | 21 |
| SELF\_CON |  | | PCL | | | |
| Mean | 33.85714286 | | Mean | | 22 | |
| Standard Error | 1.23497776 | | Standard Error | | 1.380131119 | |
| Median | 32 | | Median | | 20 | |
| Mode | 32 | | Mode | | 20 | |
| Standard Deviation | 5.659379067 | | Standard Deviation | | 6.32455532 | |
| Sample Variance | 32.02857143 | | Sample Variance | | 40 | |
| Kurtosis | 12.5751744 | | Kurtosis | | 1.959578947 | |
| Skewness | 3.262094639 | | Skewness | | 0.272621621 | |
| Range | 29 | | Range | | 28 | |
| Minimum | 27 | | Minimum | | 6 | |
| Maximum | 56 | | Maximum | | 34 | |
| Sum | 711 | | Sum | | 462 | |
| Count | 21 | | Count | | 21 | |

Now, we do the 95% confidence intervals for above 5 variables.

At 95% confidence with df=21-1=20, the critical value is 2.086 from t table.

For AGE, 95% confidence intervals for mean age is [33.81-2.086\*11.677/sqrt(21), 33.81+2.086\*11.677/sqrt(21)]=[28.495, 39.125].

For Education, 95% confidence intervals for mean age is [11.19-2.086\*2.136/sqrt(21), 11.19+2.086\*2.136/sqrt(21)]=[10.218, 12.162].

For ARR, 95% confidence intervals for mean age is [5.333-2.086\*3.088/sqrt(21), 5.333+2.086\*3.088/sqrt(21)]=[3.927, 6.739].

For SELF\_CON, 95% confidence intervals for mean age is [33.857-2.086\*5.659/sqrt(21), 33.857+2.086\*5.659/sqrt(21)]=[31.281, 36.433].

For PCL, 95% confidence intervals for mean age is [22-2.086\*6.325/sqrt(21), 22+2.086\*6.325/sqrt(21)]=[19.121, 24.879].

1. To estimate the age, race, and gender of arrestees taken into custody by a major city police department, a colleague suggests looking at data from the next 100 arrestees taken into custody beginning on Saturday, September 3, 2011. From his statistics course, he knows that a sample size of 100 is generally sufficient. Explain why this is not a good strategy. Suggest an acceptable one.

Answer: If you take the next 100 arrestees, it is called the convenience sampling. Although the sample is very easy to get, it is a biased sampling method. It is not a good strategy because there is no guarantee that the study group you have will hold a mix of natures. For example, it is possible that the next 100 arrestees are all males. And it is wrong to conclude that arrestees are all males. Hence, the group you choose is not representative of the population.

1. 3. Using data from a random sample of 100 convicted defendants, you find that the proportion of men convicted for Driving under the Influence 0.33. (The proportion of those convicted for other offenses is 0.67.) Calculate a 95% confidence interval for the proportion and explain why it is more useful than a point estimate. Is this result consistent with a null hypothesis that the true population is 0.25 (alpha = .05).

Answer: now phat=0.33. At 95% confidence, the critical value is 1.96 from standard normal sthe proportion is [0.33-0.0922, 0.33+0.0922]=[0.2378, 0.4222]

The confidence interval is more useful than a point estimate because it provides more information. Confidence intervals are constructed using a procedure that will contain the population parameter. There are infinite numbers which lie in that interval but point estimate is only one point.

Now we need to test Ho: p=0.25. At 95% confidence, the critical value is 1.96 from standard normal table. The standard deviation=sqrt(0.25\*0.75/100)=0.0433. So the test statistic=(0.33-0.25)/0.0433=1.85. So P-value=2\*P(Z>1.85)=0.0643 from standard normal table. Since p-value>0.05, we fail to reject Ho.

1. Explain the difference between the information given by a test of statistical significance, such as Chi Square and a measure of association such as lambda or gamma. Can a test of significance show a significant relationship while a measure of association shows a weak relationship between the variable? Explain your answer.

Answer: the chi-square tests the probability that there is more than a random association between two variables. However, the measure of association examines the association between two variables regardless of the result of chi square test.

Yes, a test of significance can show a significant relationship while a measure of association shows a weak relationship between the variables because there are limitations of chi square such as sample size. For example, for small size of sample, the chi square test will report that there is no association between two variables but the measure of association is quite large.