1. Multiple-choice section

1-1 Business school

A business school claims that only 15% of its students are still unemployed one year after graduation. You are suspicious the school may not be living up to its claim (i.e., that the actual unemployment rate is higher) and have collected a random sample of 130 students from the school that graduated a little over a year ago. You find that 24 of them have not yet found a job. What is the p-value of the test for the school’s claim? (Hint: use the parameter testing.xlsx template).

A. 0,1345

B. 0,1846

C. 0,2690

D. 0,8655

Answer:

1-2 Brexit

A political scientist wishes to estimate the proportion of British people who are still in favor of leaving the EU. How large a sample should she collect to estimate this proportion with an error margin of no more than 2 percentage points (at 95% confidence)?

A. 2401

B. 626

C. 102

D. There is not enough information to tell.

Answer:

1-3 Interval vs. point estimate

An interval estimate is typically preferred over a point estimate because …

A. it gives us the probability that it contains the true parameter (e.g., 95%).

B. it provides us with more possible parameter values.

C. it gives us a sense of the accuracy of the point estimate.

D. All of the above.

Answer:

1-4 Margin of error

A 95% confidence interval for the average hours of TV watching by residents in Frederiksberg Kommune, based on a random sample of 200 residents is an example of…

A. a parameter.

B. a summary measure.

C. descriptive statistics.

D. inferential statistics.

Answer:

1-5 Graphing

Parking in central Copenhagen has become a major issue. City planners would like to determine the average time it takes a person to find a parking spot from the time they arrive in the City Center. Drivers participating in the study were asked to record the time between entering the center and pulling into a parking spot. Which of the following would **not** be appropriate for displaying the search time data?

A. Histogram

B. Box plot

C. Pie chart

D. None of these should be used

Answer:

1-6 Tennis serves

At a tennis tournament a statistician keeps track of every serve that a player hits. He finds that the mean serve speed of a particular player is 158 km/h and that the distribution of serve speeds is skewed to the left. Which of the following values might then be the median serve speed?

A. 142 km/h

B. 150 km/h

C. 158 km/h

D. 166 km/h

Answer:

1-7 Sensitive to outliers

Which of the following numerical summary measures is **NOT** sensitive to outliers in a dataset?

A. Mean

B. Standard deviation

C. Range

D. Inter-quartile range

Answer:

1-8 Correlation coefficient

Which of the following statements about the correlation coefficient r between X and Y is **NOT** correct?

A. r does not depend upon the units of X and Y.

B. r does not depend upon which variable is treated as the response variable.

C. r is always between -1 and +1.

D. r measures the strength of any kind of relationship between X and Y.

Answer:

1-9 Lurking variable

A study shows that the amount of chocolate consumed in Mexico City and the number of automobile accidents is positively related. Which of the following could be a potential lurking variable?

A. Vehicle speeds

B. Urban infrastructure

C. Population growth

D. No lurking variable.

Answer:

1-10 Sample frame

Ideally, the sample frame includes …

A. a simple random sample.

B. a convenience sample.

C. the entire sample.

D. the entire population of interest.

Answer:

1-11 Probabilities

For two events and , we are told that and and that the two events are independent. Then the probability of (“ and/or ”) equals

A. .

B. .

C. .

D. .

Answer:

1-12 Sampling distribution

The closer the true proportion is to either 0 or 1, the larger the sample size must be in order for the sampling distribution of the sample proportion to …

A. be approximately normal (bell-shaped).

B. be unbiased, i.e. centered at the true value .

C. be a consistent estimate of .

D. have a standard deviation of .

Answer:

1-13 Multiple regression equation

In a multiple regression problem involving two explanatory variables , if the coefficient is estimated to be , it means that …

A. the relationship between and is significant.

B. the predicted value will increase by units if we increase by 1 unit and hold constant.

C. the predicted value will increase by units if we increase by 1 unit whether or not we make any changes in

D. the predicted value when .

Answer:

1-14 P-value

A large p-value indicates that …

A. there is no evidence against the null hypothesis

B. the null hypothesis is most likely true

C. the alternative hypothesis is not plausible

D. the alternative hypothesis is plausible

Answer:

1-15 Dependent sampling

The main advantage of dependent (paired) sampling over independent sampling to compare two conditions is that …

A. it requires fewer subjects and is therefore cheaper.

B. it allows one to eliminate the variation between pairs and thus focus more on the difference between the two conditions for each pair.

C. it leads to a larger difference between the two conditions

D. it creates more variation between pairs, allowing for a wider range of observations.

Answer:

2. Analytical section

Question 2-1 (Scandinavian market study)

You are doing a market study for a company that operates in the four Scandinavian countries (Denmark, Norway, Sweden, Finland) and want to know how certain factors, most importantly price, affect their sales.

You have collected information from 100 locations, approximately an equal number in each country, and recorded both the sales in that location (Sales), the average price of the product (Price), the product quality, as assessed by the local customers as an index between 0 and 1 (Quality), the customer satisfaction of the local customers as index between 0 and 1 (Satisfaction), and finally what country the location is situated in (Country). The data are found in the JMP file “Question 2-1.JMP”.

Use JMP to perform a multiple regression of Sales as a function of price, quality, satisfaction, and country. In the interest of time, you are not required to look at residuals or distributions of variables. Then answer the following questions. Please include output from JMP and refer to it in your answers.

Question 2-1-a

Is the regression model as a whole able to “explain” or predict sales? And “how well” does it predict sales?

Answer:

Question 2-1-b

Comment on the estimated coefficients in the regression. Are they significant? Do they have the expected sign? Does country have any influence on sales?

Answer:

Question 2-1-c

Use the model to predict sales in a Norwegian town where the price is 5,0, the perceived quality is 0,6, and customer satisfaction is 0,6. (Hint: You may want to choose “Estimates/Show prediction equation” in the dropdown menu in the regression output window to help you out).

Answer:

Question 2-1-d

Consider those coefficients that were not significant in your regression. Could the lack of significance be due to a problem in the data? Use an appropriate analysis method to see whether your hunch is correct. (Hint: consider multicollinearity.)

Answer:

Question 2-1-e

Modify (simplify) your model to avoid the problem found in (d). Compare the results to the previous analysis: Is this more in line of what you would expect to see?

Answer:

Question 2-2 (Soft drink market)

A company has conducted a study of the Danish market for a new soft drink “Nu-Drink”, geared towards young persons between the age of 18 and 25. Using appropriate methods, they have taken a random sample of 200 young Danes in that age group and let them compare “Nu-Drink” to the three competing brands, “7-up”, “Sprite”, and “Mountain Dew” and recorded the following information:

* What region they are from (REGION)
* What brand of soft drink they preferred in the test (DRINK)
* How much they typically spend per year on soft-drinks (SPEND)

The regions in the study are: Copenhagen and environs: “Cph”, Århus-Ålborg-Odense: “City”, Northern and Middle Jutland: “Jutland”, and the rest of the country: “Rest”.

The data is found in the JMP table “Question 2-2.JMP”. If you need it, you can also find the same data in the Excel file “Question 2-2.xlsx”. At this point in your project, you are primarily interested in whether there might be a difference between the different regions. In the interest of time, you do not need to inspect the data for outliers, errors, or skewed distributions.

Question 2-2-a

Create a graph that illustrates a possible difference between the regions in terms of what brand of soft drink people prefer. (You can either use an Excel pivot-chart to do this or use the “Mosaic plot” in the output from JMP in your answer to the next question below.)

Answer:

Question 2-2-b

Conduct a formal statistical test to check whether there is a difference in (a). If you use 5% as your significance level, what do you conclude?

Answer:

Question 2-2-c

Create a graph that illustrates a possible difference between the regions in terms of how much people spend on soft drinks per year. (You can either use Excel to do this or use the output from JMP in your answer to the next question. The latter is probably by far the easiest.).

Answer:

Question 2-2-d

Conduct a formal statistical test to check whether there is a difference in between then regions in (c). If you use 5% as your significance level, what do you conclude?

Answer:

3. Discussion section

3-1 False positive results in statistics

In recent years, there have been an increasing number of warnings from statisticians that researchers rely too much on the p-value and the famous as a criterion for considering results to be significant.

3-1-a

In theory, implies that there is up to a 5% chance of getting a false positive results (type 1 error). Briefly explain (in half a page or less) how one could eliminate or minimize false positive results (type 1 errors) and why in practice this rarely happens so that such errors tend to persist.

Answer:

3-1-b

Briefly explain (in half a page or less) how the practice of statistics data collection and modeling can lead to type 1 errors much more frequently than the 5% suggested by a significance level of .

Answer:

3-1-c

What are some principles or guidelines for doing research that can help minimize the problems discussed in 3-1-a and 3-1-b?

Answer: