1. Intergalactic scientists recruited by the Rebellion were tasked with studying midi-chlorian levels in children and young adults (aged 5 to 25). Midi-chlorian counts are linked to sensitivity to The Force and may confer extraordinary abilities. Normal “Human” midi-chlorian levels average approximately 2,500 per cell and much higher levels of midi-chlorian are observed in Jedi. Using a large and random sample, these scientists determine that the standard deviation of mid-chlorian counts is approximately 625 per cell. These Rebel Scientists wish to identify individuals with sufficient Force sensitivity (i.e., potential to be trained as new Jedi) and have declared that for now, only Humans with a z-score of at least 2.4 may qualify.

a) What is the minimum mid-chlorian count that would be classified as being sufficiently Force-sensitive?

b) What proportion of the Human population can potentially be trained as new Jedi?

c) Midi-chlorian counts lower than 2,000 provide no sensitivity to the Force. What proportion of the Human population are therefore useless to train for the Jedi Order?

d) What assumptions must the Rebels make for these data to be accurate?

2. A new study published in the esteemed *New Republic Journal of Medicine* reported results from a randomized trial designed to assess the effect of a drug called **bota** on boosting the effects of midi-chlorians to enhance a Human’s Force-sensitivity. The study reports a model including bota as a binary variable called **Exposure** that indicates treatment with the drug (0 = no; 1 = yes), gender as a binary variable called **Female** (0 = male; 1 = female), age in years, and the number of servings of fruits consumed per day. The dependent variable is defined as enhanced Force-sensitivity (0 = no; 1 = yes).

Table

Description automatically generateda) Write the complete regression equation for the model provided above.

b) Interpret the number in the “odds ratio” column for the variable **Exposure** in one sentence.

c) Interpret the number in the “odds ratio” column for the variable **Fruits** in one sentence. d) The value represented by the letter **A** is (please **bold** / highlight / or somehow indicate your answer choice):

Less than 0

Greater than 0

This information cannot be determined from the data provided

The value represented by the letter **B** is (please **bold** / highlight / or somehow indicate your answer choice):

Less than 0

Greater than 0

This information cannot be determined from the data provided

Do you think that the Rebel Scientists should give bota to Humans to enhance their Force-sensitivity? In no more than two sentences, report why or why not.

3. Given the paucity of Force-sensitive Humans, Rebel Scientists begin to consider environmental factors that might be associated with diminished midi-chlorian counts. Several chemicals were evaluated. There was particular interest in assessing serum levels of the element Fermium, a heavy metal that does not occur naturally. Below are linear regression output for fermium followed by questions.

Table

Description automatically generated

a) Quantify the relationship between fermium and midi-chlorian count in one sentence.

b) What can you say, if anything, about the correlation coefficient between serum Fermium and midi-chlorian count?

c) If a randomly selected Human has a serum Fermium level 10 units lower than their neighbor, what is the best prediction for their difference in midi-chlorian count?

d) Interpret the \_cons term.

Although not shown, you are told that the R2 for the model above is 0.17.

e) What does this mean?

f) Given this information, how would you advise the Rebel Scientists with respect to next steps?

4. Suppose that the Rebel Scientists further adjusted their simple model from the previous question (i.e., the model in Question 3 above assessing the association between serum Fermium and midi-chlorian counts) for age, gender, and family history of Force-sensitivity (0 = no; 1 = yes).

a) Write the full multivariable regression model.

b) They found that the coefficient for “family history” was 10.26 (95% confidence interval: 7.13 to 13.38, p = 0.0003). Interpret the slope and define the p-value and 95%CI.

c) The R2 for the new multivariable model is 0.57. In no more than two sentences: 1) interpret this statistic; and 2) comment on the difference from the R2 of 0.17 that was reported for the simple model.

d) Imagine that you asked the Scientists why they decided to focus on Fermium, and the team replied that they simply tested every element from the periodic table to see what was related to midi-chlorian counts. In no more than two sentences, tell me: 1) what statistical topic would immediately come to your mind; and 2) what would you say to the Rebel Scientists?

5. One Rebel Scientist shows you this output from a multiple linear regression. She investigated how age affects the levels of midi-chlorian in the blood. She suspected potential effect modification by gender with respect to the outcome, and therefore appropriately included an interaction term (**inter**) in her model. The interaction term was constructed by multiplying the **age** variable by the **Female** variable (0 = male; 1 = female).

a) Write the complete regression equation for the model provided below.

b) Write separate regression equations for men and women.

c) In one sentence, **using no numbers** (i.e., in simple English), provide an interpretation of the statistically significant effect modifier with respect to the variables under study.

d) The Rebel Scientist tells you that based on this output, female Humans have midi-chlorian counts that are 40 units higher than males after adjusting for age. In one sentence, what would you say about this interpretation of her results?

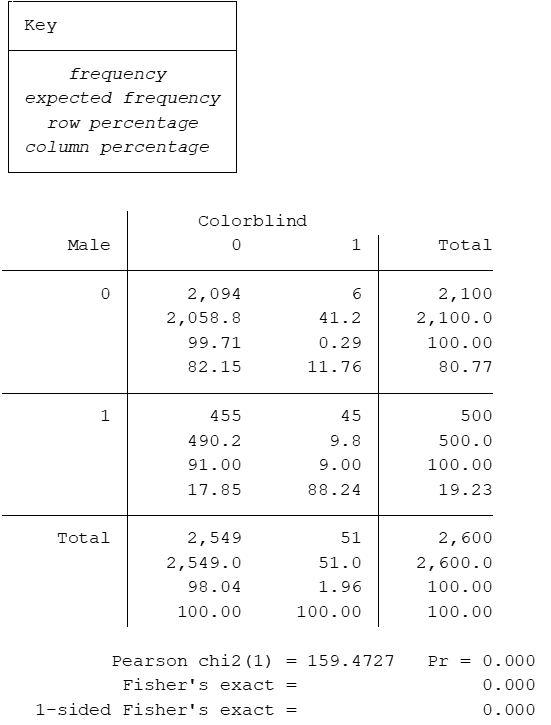
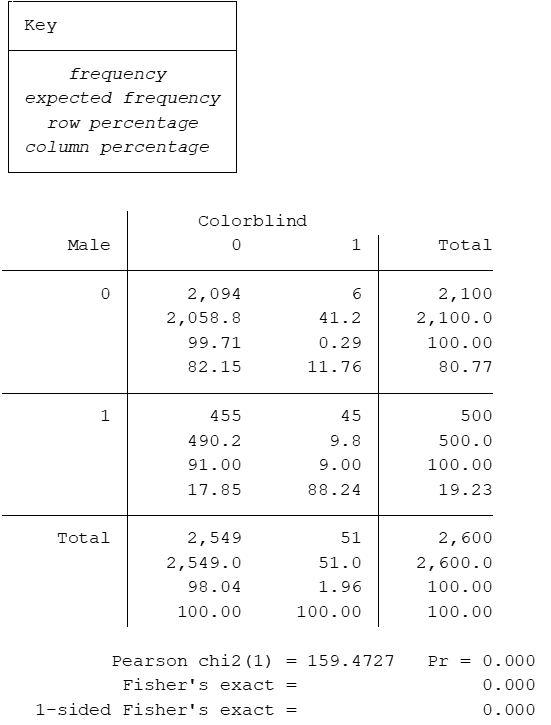
Table

Description automatically generated

6. Jedi Knights have lightsabers of varying colors, including green, blue, and even purple. The Sith always have red lightsabers. The Jedi Counsel therefore believes it might be prudent to conduct a study of red / green color blindness. A total of 500 male and 2100 female Jedi are randomly selected and tested for color blindness. Data are coded such that “1” indicates the presence of a characteristic and “0” indicates its absence. Use the data at the bottom to answer the following questions:

a) Based on the contingency table provided, what research question is the Counsel aiming to address?

b) In one sentence, summarize the findings of these data.



7. For the following research scenarios, state the most appropriate statistical method that could be used to answer each question. **Please carefully consider the nature of the variables given.** If you feel that there is not enough information provided (including ambiguous coding for a given variable) and additional information would indicate use of a particular test, please state the assumption you are making.

Jedi Master Yoda recommended that Force-sensitive individuals over a certain age should **not** be trained as Jedi; however, we know that some Jedi (e.g., Anakin Skywalker) began their training at older ages. Is age when training began associated with whether or not a Jedi turns to the Dark Side?

Older Jedi Padawans (trainees) often have personal attachments that might be related to turning to the Dark Side. Is having personal attachments (yes / no) associated with turning to the Dark Side?

You are curious about the mean midi-chlorian count of Wookiees, Gungans, and Ewoks. How could you test whether midi-chlorian count differs between these species?

An expert statistician points out that the midi-chlorian count in Wookiees, Gungans, and Ewoks might not be normally distributed, and the median might be a better measure of central tendency than the mean. How can you test whether the median midi-chlorian count differs between these species?

Does the average smuggler make the Kessel Run in 12 parsecs?

Among starships with hyperdrives (i.e., the equipment that allows a ship to travel at the speed of light), is there a relationship between eco-friendliness of the ship (*terrible*, *mediocre*, *good*, or *fantastic*) and the average number of times per month the hyperdrive is used?

After controlling for a variety of variables, does the number of sandstorms per year predict the average annual temperature on the desert planet of Tatooine?

Are children with a family history of Jedi training more likely to become Jedi themselves?

Is the average luminescent blade of a lightsaber actually 3 feet long?

Is political loyalty (Rebel / Empire) associated with Starship preference (X-wing / TIE fighter / Imperial Star Destroyer / Rebel Transport)?

8. This question will require you to analyze some data in Stata. However, all data related to the Jedi are proprietary, highly classified, and are not publicly available. We therefore must (sadly) return to the publicly available CALERIE trial for the final component of this final. Recall that the NIH-funded CALERIE trial (**C**omprehensive **A**ssessment of **L**ong-term **E**ffects of **R**educing **I**ntake of **E**nergy; <https://calerie.duke.edu/>) is a randomized controlled trial that aimed to determine the effect of a calorie restriction (CR) intervention on markers of longevity, disease risk, and quality of life in healthy adults without obesity. Please use the dataset ***calerieagain.dta*** to answer the following questions. A codebook for this dataset is included on the last page of the final

**Context**: Intentional calorie restriction (CR) results in weight loss, even among individuals without obesity. We would therefore like to determine whether randomization to a CR versus an *ad libitum* Control intervention is associated with changes in weight over 12 months.

Conduct an appropriate statistical test to test whether there is an association between being randomized to the treatment or control group (*treatment*) and the percentage of baseline body weight lost at 12 months (*pctwt*). Interpret the coefficient for *treatment*. Be mindful of how to interpret the directionality of the variable *pctwt*.

There are likely other factors that contribute to weight loss in addition to an intervention. Conduct an appropriate statistical test to quantify the association between randomization (*treatment*) and weight loss (*pctwt*) after adjustment for energy intake (*kcal*), emotional wellbeing (*ewbscore*), and education (*educate*). Interpret the coefficient for each of these variables.

Quantify the accuracy of the prediction for percentage weight change in one sentence.

Interpret the R2 value for the multivariable model in one sentence. Comment on whether the multivariable model is preferable to the results you observed in (A).

Conduct two diagnostic tests (one of which must be standardized residuals versus standardized predicted graph) to assess how well your model fits the data. Briefly assess your model fit.

Create a well-designed table (i.e., a Table 1) **in** **Microsoft Word** that gives a reader an overview of how the CR and Control groups look at baseline with respect to the following variables: age, sex, race, education, weight, energy intake, percentage of energy consumed from fat, protein, and carbohydrate, general health, and emotional wellbeing.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Coding scheme / units** |
| id | deidentified subject number | integer |
| AGEBL | age at baseline visit | years |
| GENDER | sex | 1 = male 2 = female |
| RACE | race | 2 = Asian 4 = Black or African American 5 = White 8 = Other |
| EDUCATE | education | 0 = Less a college degree 1 = College degree 2 = More than a college degree |
| MCLINWT | mean clinic weight | kilograms (kg) |
| KCAL | self-reported energy intake | kilocalories/day (kcal/d) |
| PCTFATC | self-reported % calories from fat | percentage per week (%) |
| PCTPROTC | self-reported % calories from protein | percentage per week (%) |
| PCTCARBC | self-reported % calories from carbs | percentage per week (%) |
| EWBSCORE | Health-related quality of life: emotional wellbeing | score |
| GHSCORE | Health-related quality of life: general | score |
| pctwt | Percent weight change from baseline to 12 months, calculated as ((weight12 - weight0)/(weight0))\*100 | percentage over 12 months (%) |
| treatment | randomized treatment arm | 1 = caloric restriction 0 = ad libitum control |