1. Does your Zodiac sign predict how successful you will be later in life? Fortune magazine collected the zodiac signs of 256 CEOs of the largest 400 companies. The table shows the number of births for each sign. Are the zodiac signs of CEOs distributed uniformly?

State: 
\text{H}_0: \text{Births are uniformly distributed over zodiac signs} \\
\text{H}_a: \text{Births are not uniformly distributed over zodiac signs}

Do: \text{df = 11} \\
\chi^2 = \frac{(23-21.33)^2}{21.33} + \frac{(20-21.33)^2}{21.33} + \ldots + \frac{(29-21.33)^2}{21.33} = 5.094 \\
P\text{-value} = \chi^2\text{cdf}(5.094, \infty, 11) = 0.926

Conclude: Because our P-value 0.926 > 0.05, we do not reject \text{H}_0. We do not have convincing evidence that the births are not uniformly distributed.

2. Many colleges survey graduating classes to determine the plans of the graduates. We might wonder whether the plans of students are the same at different colleges. Here’s a two-way table for Class of 2006 graduates from several colleges. Are the student’s choices of post-graduation activities the same across all the colleges?

<table>
<thead>
<tr>
<th></th>
<th>Agriculture</th>
<th>Arts and Science</th>
<th>Engineering</th>
<th>Social Studies</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employed</td>
<td>379</td>
<td>305</td>
<td>243</td>
<td>125</td>
<td>1052</td>
</tr>
<tr>
<td>Grad School</td>
<td>186</td>
<td>238</td>
<td>202</td>
<td>96</td>
<td>722</td>
</tr>
<tr>
<td>Other</td>
<td>104</td>
<td>123</td>
<td>37</td>
<td>58</td>
<td>322</td>
</tr>
<tr>
<td>Total</td>
<td>669</td>
<td>666</td>
<td>482</td>
<td>279</td>
<td>2096</td>
</tr>
</tbody>
</table>

State: 
\text{H}_0: \text{Students post grad activities are distributed the same for all four colleges} \\
\text{H}_a: \text{Students post grad activities are not distributed the same for all four colleges}

Do: \text{df = (3-1)(4-1) = 6} \\
\chi^2 = \frac{(379-335.77)^2}{335.77} + \ldots + \frac{(58-42.86)^2}{42.86} = 54.52 \\
P\text{-value} = \chi^2\text{cdf}(54.52, \infty, 6) = 0.000000006

Conclude: Because our P-value < 0.005, we reject \text{H}_0. We have convincing evidence that students’ postgrad activities are not evenly distributed for each of the colleges.
3. A study from the University of Texas Southwestern Medical Center examined whether the risk of hepatitis C was related to whether people had tattoos and to where they got their tattoos. Hepatitis C causes about 10,000 deaths each year in the United States, but often goes undetected for years after infection. The data from this study can be summarized in a two-way table, as follows:

<table>
<thead>
<tr>
<th></th>
<th>Hepatitis C</th>
<th>No Hepatitis C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tattoo, Parlor</td>
<td>17</td>
<td>35</td>
<td>52</td>
</tr>
<tr>
<td>Tattoo, Elsewhere</td>
<td>8</td>
<td>53</td>
<td>61</td>
</tr>
<tr>
<td>None</td>
<td>22</td>
<td>491</td>
<td>513</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>579</td>
<td>626</td>
</tr>
</tbody>
</table>

Are “tattoo status” and “hepatitis status” independent?

**State:**

- $H_0$: There is no association between tattoo status and hepatitis status
- $H_a$: There is an association between tattoo status and hepatitis status

- $\alpha=0.05$

**Plan:** Chi-square test for independence

- Random: Assume the sample was randomly selected
- 10% Cond: 626 patients are fewer than 10% of all patients with tattoos
- X Large Counts: Not all expected counts $\geq 5$

**Do:**

- $df = (3-1)(2-1) = 2$
- $\chi^2 = \frac{(17-3.904)^2}{3.904} + \frac{(35-48.906)^2}{48.906} + \ldots + \frac{(491-474.8)^2}{474.8} = 57.91$
- $p-value = \chi^2 df (57.91, 2, 0.0001)$

**Conclude:**

Because our $p$-value $0.0001 < \alpha=0.05$, we reject $H_0$. There is convincing evidence of an association between hepatitis status and tattoo status. Because the “Large Counts” was violated, we need to check that the two cells with small expected counts did not influence this result too greatly.
4. A high-school administrator who is concerned about the amount of sleep the students in his district are getting selects a random sample of 14 seniors in his district and asks them how many hours of sleep they get on a typical school night. He then uses school records to determine the most recent grade point average (GPA) for each student. His data and computer regression output are given below.

<table>
<thead>
<tr>
<th>Sleep (hrs)</th>
<th>9</th>
<th>8.5</th>
<th>9</th>
<th>7</th>
<th>7.5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>5.5</th>
<th>6</th>
<th>8.5</th>
<th>8</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>3.8</td>
<td>3.3</td>
<td>3.5</td>
<td>3.6</td>
<td>3.4</td>
<td>3.3</td>
<td>3.2</td>
<td>3.2</td>
<td>3.4</td>
<td>3.4</td>
<td>3.6</td>
<td>3.1</td>
<td>3.4</td>
</tr>
</tbody>
</table>

Predictor Coef SE Coef T  P
Constant 2.6476 0.3281 8.07 0.000
Sleep 0.10197 0.04347 2.34 0.037

- S = 0.180471 R-Sq = 31.3% R-Sq(adj) = 25.6%

a. Do these data provide convincing evidence of a linear relationship between hours of sleep students typically get and their academic performance, as measured by their GPA? Carry out a significance test at the 0.05 significance level.

State: $H_0: \beta = 0$  \quad $\beta$ = True slope of the population regression line relating hrs of sleep to GPA for seniors at this high school.

Do: $df = 14 - 2 = 12$

$\hat{t} = \frac{0.10197 - 0}{0.04347} = 2.34$

$p\text{-value} = \text{tcdf}(2.34, 12) = 0.01869$

Conclude: Since our $p\text{-value} 0.01869 < \alpha = 0.05$, we reject $H_0$. We have convincing evidence that there is a linear relationship between hrs of sleep and GPA.

b. Construct and interpret a 95% confidence interval for the slope of the regression of GPA on hours of sleep for seniors in this school district.

State: Estimate $\beta$ with 95% confidence

Do: $df = 12$

$\text{InvT}(0.025, 12) = 2.179$

$0.10197 \pm 2.179(0.04347)$

$0.0071, 0.1965$

Conclude: We are 95% confident the interval $(0.0071, 0.1965)$ captures the true slope of the population regression line relating GPA to hrs of sleep for seniors in this district.

c. Can we conclude from these data that students’ GPA will improve if they get more sleep? Explain.

No because this is an observational study and not a controlled experiment. There may be other factors that are confounded with the amount of sleep students get.
5. An important advance in ecological research from the late 1960’s was the Theory of Island Biogeography, which demonstrated a relationship between the size of islands and the number of species of different groups that could be found in the islands. One of these first examples found was the number of amphibian and reptile species on islands in the West Indies. The data and a scatterplot of number of species versus island area (in square miles) is given below.

<table>
<thead>
<tr>
<th>Island</th>
<th>Area (sq. miles)</th>
<th>Number of species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redonda</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Saba</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Montserrat</td>
<td>104</td>
<td>9</td>
</tr>
<tr>
<td>Puerto Rico</td>
<td>8816</td>
<td>40</td>
</tr>
<tr>
<td>Jamaica</td>
<td>11669</td>
<td>39</td>
</tr>
<tr>
<td>Hispaniola</td>
<td>77793</td>
<td>84</td>
</tr>
<tr>
<td>Cuba</td>
<td>103723</td>
<td>76</td>
</tr>
</tbody>
</table>

a. Clearly, this relationship is not linear. Find a logarithmic transformation that produces a clearly linear relationship. Support your choice with appropriate graphical or numerical analysis of the data.

\[
\log(\text{Area}) = 0.3307 + 0.3165 \log(\text{number of spec})
\]

\[
\hat{r} = 0.9987
\]

\[
\hat{r}^2 = 0.9975
\]

The scatterplot is linear and the residual plot shows no obvious pattern.

b. Use your model to predict the number of amphibian and reptiles species on the island of Martinique, which has an area of 436 square miles.

\[
\log(\text{Area}) = 0.3307 + 0.3165 \log(436)
\]

\[
\log(\text{Area}) = 1.1661
\]

\[
10^{1.1661} = \hat{\text{Area}}
\]

\[
14.66 = \hat{\text{Area}}
\]