6.34 Barking deer. Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests make up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.

Woods	Cultivated grassplot	Deciduous forests	Other	Total
4	16	61	345	426

(a) Write the hypothesis for testing if barking deer prefer to forage in certain habitats over tohers.

H_0 : The proportion of barking deer in each habitat is equal.

H_1 : The proportion of barking deer in each habitat is not equal.

(b) What type of test can we use to answer this research question?

A χ-squared test can be used to answer this research question.

(c) Check if the assumptions and conditions required for this test are satisfied.

The expectation value of each category is ≥ 5.

(d) Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

Using the hypothesis stated above the test statistic is given by

$$
\begin{aligned}
X^2 &= \sum \frac{(\text{observed} - \text{expected}^2)}{\text{expected}}, \\
&= \frac{(4 - 0.048 \cdot 426)^2}{0.048 \cdot 426} + \frac{(16 - 0.147 \cdot 426)^2}{0.147 \cdot 426} + \frac{(61 - 0.396 \cdot 426)^2}{0.396 \cdot 426} \\
&\quad + \frac{(345 - (1 - 0.048 - 0.147 - 0.396) \cdot 426)}{(1 - 0.048 - 0.147 - 0.396) \cdot 426}, \\
&= 248.889.
\end{aligned}
$$

The degrees of freedom (dof) is given by $k - 1$ where $k = 4$ is the number of categories. The p-value is given by

$$
\begin{aligned}
\text{p-val} &= P(\chi^2 \geq X^2), \\
&= P(\chi^2 \geq 248.889), \\
&\overset{R}{=} 1 - \text{pchisq}(248.889, 3), \\
&\overset{R}{=} 0.
\end{aligned}
$$

For $\alpha = 0.05$, p-val $< \alpha$. Thus, we reject the null and accept the alternative: barking deer do not forage in all habitats equally.

6.50 Coffee and Depression. Researchers conducted a study invetigating the relationship between cafeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The reseaerchers used qesionnaires to collect dta on caffeinated coffee consumption, asked each individual about physician- diagnosed depression, and also asked about the use of antidepressans. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.

		Caffeinated coffee consumption					
		≤ 1	2 to 6	1	2 to 3	≥ 4	
		cups/week	cups/week	cups/day	cups/day	cups/day	Total
Clinical	Yes	670	373	905	564	95	2607
depression	No	11545	6244	16329	11726	2288	48132
	Total	12215	6617	17234	12290	2383	50739

(a) What type of test is appropriate for evaluating if there is an association between cofee intake and depression?

A χ-squared test for two-way table.

(b) Write the hyothesis for the test you identified in part (a).

H_0 : There is no association.

H_1 : There is some association.

(c) Calculate the overall proportion of women who do and do not suffer from depression.

Let p_s be the proportion of women who suffer from depression and p_n be the proportion of women who do not.

$$p_s = \frac{2607}{50739}, \quad p_n = \frac{48132}{50739},$$
$$p_s = 0.0514, \quad p_n = 0.949.$$

(d) Identify the expected count for the highlighted cell (2 to 6 cups/week, yes), and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2 / Expected$.

Let $E_{\text{2 to 6, yes}}$ be the expected value for the highlighted cell. Then

$$\begin{aligned} E_{\text{2 to 6, yes}} &= \frac{2607}{5}, \\ &= 521.4. \end{aligned}$$

The contribution to the test statistic is given by

$$\begin{aligned} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} &= \frac{(373 - 521.4)^2}{521.4}, \\ &= 42.24. \end{aligned}$$

(e) The test statistic is $\chi^2 = 20.93$. What is the p-value?
The dof are

$$\begin{aligned} \text{dof} &= (\#\text{ of rows} - 1)(\#\text{ of cols} - 1), \\ &= (2 - 1)(5 - 1), \\ &= 4. \end{aligned}$$

The p-value is given by

$$\begin{aligned} \text{p-value} &= P(X^2 \geq \chi^2), \\ &= P(X^2 \geq 20.93), \\ &\overset{R}{=} 1 - \text{pchisq}(20.93, \text{dof} = 4), \\ &\overset{R}{=} 0.0003269507. \end{aligned}$$

(f) What is the conclusion of the hypothesis test?
For $\alpha = 0.05$, p-value $< \alpha$. Thus, we reject the null and accept the alternative: there is an association between caffeine consumption and depression in women.

(g) One of the authors of this study was quoted on the NYTimes as saying it was “too early to recommend that women load up on extra coffee” based on just this study. Do you agree with this statement? Explain your reasoning.
Yes, it is to early to conclude that women should drink coffee to prevent depression. The hypothesis test we conducted only demonstrated a relationship, not that coffee prevents depression. Furhtermore this only demonstrates an association, not a causal relationship.

7.6 Working backwards, Part II. A 90% confidence interval for a population mean is $(65, 77)$. The population distribution is approximately normal and the population standard deviation is unkown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation.

The confiddence interval is given by

$$\mu \pm t^* \cdot \frac{s}{\sqrt{n}},$$

where μ is the sample mean, t^* is given by R using the confidence level and dof, s is the sample standard deviation, and $n = 25$ is the sample size. The margin of error, $\neptune = t^* \frac{s}{\sqrt{n}}$. Then

$$\begin{aligned} t^* &\overset{R}{=} \mathrm{qt}(0.9, n-1), \\ &\overset{R}{=} \mathrm{qt}(0.9, 24), \\ &\overset{R}{=} 1.710882. \end{aligned}$$

Thus

$$77 = \mu + \neptune = \mu + t^* \frac{s}{\sqrt{n}}, \quad 65 = \mu - \neptune = \mu - t^* \frac{s}{\sqrt{n}}.$$

Thusly

$$\begin{aligned} 77 - 65 &= \mu + \neptune - (\mu - \neptune), \\ 12 &= 2\neptune, \\ 6 &= \neptune. \end{aligned}$$

Thusmore

$$\begin{aligned} 77 &= \mu + \neptune, \\ 77 - \neptune &= \mu, \\ 71 &= \mu. \end{aligned}$$

Thusmoreover

$$\begin{aligned} \neptune &= t^* \frac{s}{\sqrt{n}}, \\ \frac{\neptune\sqrt{n}}{t^*} &= s, \\ \frac{6\sqrt{25}}{1.710882} &= s, \\ 17.53 &= s. \end{aligned}$$

7.12 Auto exhaust and lead exposure. Researchers interested in lead exposure due to car exhaust sampled the blood of 52 police officers subjected to constant inhalation of automobile exhaust fumes while working traffic enforcement in a primarily urban environment. The blood samples of these officers had an average lead concentration of $124.32\,\mu$g/l and a SD of $37.74\,\mu$g/l; a previous study of individuals from a nearby suburb, with no history of exposure, found an average blood level concentration of $35\,\mu$g/l.

(a) Write down the hypothesis that would be appropriate for testing if the police officers appear to have been exposed to a different concentration of lead.

The null and alternative hypothesis are

H_0 : The officers were exposed to the same concentration of lead.

H_1 : The officers were exposed to a difference concentration of lead.

(b) Explicitly state and check all conditions necessary for inference on these data.

We assume a simple random sample was preformed since it is not stated otherwise. Since $n \geq 30$ the data is nearly normal barring extreme outliers. Since we are not given the actual data we will assume there are no extreme outliers.

(c) Regardless of your answers in part (b), test the hypothesis that the downtown police officers have a higher lead exposure than the group in the previous study. Interpret your results in context.

The test statistic t is given by

$$\begin{aligned} t &= \frac{124.32\frac{\mu\text{g}}{\text{l}} - 35\frac{\mu\text{g}}{\text{l}}}{\frac{s}{\sqrt{n}}}, \\ &= 18.4027. \end{aligned}$$

With dof $= 52 - 1$,

$$\begin{aligned} \text{p-val} &= 2P(T > |t|), \\ &= 2P(T > 18.4027), \\ &= 2P(T < -18.4027), \\ &\overset{R}{=} 2 * \text{pt}(-18.4027, 51), \\ &\overset{R}{=} 1.798\text{e}{-24}. \end{aligned}$$

For $\alpha = 0.05$ p-val $< \alpha$. Thus, we reject the null and accept the alternative: the officers were exposed to a difference concentration of lead.

7.14 SAT scores. The standard deviation of SAT scores for students at a particular Ivy League college is 250 points. Two statistics students, Raina and Luke, want to estiate the average SAT score of studens at this college as part of a class project. They want their margin of error to be no more than 25 points.

(a) Raine wants to use a 90% confidence interval. How large a sample should she collect?

Let n be the size of Raines sample. Then

$$\begin{aligned} 25 &\leq t^*_{0.9}\frac{s}{\sqrt{n}}, \\ \frac{1}{10} &\leq \frac{t^*_{0.9}}{\sqrt{n}}, \\ \frac{1}{10} &\overset{R}{\leq} \frac{\mathrm{qt}(0.95, n-1)}{\sqrt{n}}, \\ 273 &\overset{R}{\leq} n. \end{aligned}$$

(b) Luke wants to use a 99% confidence interval. Without calculating the actual sample size, determine whether his sample should be larger or smaller than Raina's, and explain your reasoning.

The sample size for Luke's confidence interval must be larger. Geometrically, by increasing the confidence interval width there is a greater area under the distribution and so n must increase to reduce the overall quantity.

(c) Calculate the minimum required sample size for Luke.

$$\begin{aligned} 25 &\leq t^*_{0.99}\frac{s}{\sqrt{n}}, \\ \frac{1}{10} &\overset{R}{\leq} \frac{\mathrm{qt}(0.995, n-1)}{\sqrt{n}}, \\ 668 &\overset{R}{\leq} n. \end{aligned}$$

7.22 High school and beyond, Part II. We considered the difference between the reading and writing scores of a random sample of $n = 200$ students who took the High School and Beyond Survey in Exercise 7.20. The mean and standard deviation of the differences are $\bar{x}_{\text{read-write}} = -0.545$ and $s = 8.887$ points.

(a) Calculate a 95% confidence interval for the average difference between the reading and writing scores of all students.

$$\bar{x}_{\text{read-write}} \pm t^*_{0.95} \frac{s}{\sqrt{n}} \overset{R}{=} -0.545 \pm \mathrm{qt}(0.955, 200 - 1) \frac{8.887}{200}.$$

Thus the confidence interval is

$$(-0.621, -0.469).$$

(b) Interpret this interval in context.

We are 95% confident that the true value for the difference between reading and writing scores of students who took the High School and Beyond Survey lies between -0.621 and -0.469 quantile.

(c) Does the confidence interval provide convincing evidence that there is a real difference in the average scores? Explain.

Since the confidence interval is entirely negative we are 95% confident that the mean writing score is higher than the mean reading score.