1. To determine how the calcium level of water affects respiration rates for fish, 90 fish were randomly divided into three tanks with different levels of calcium (*low*, *medium*, and *high*) and their respiration rates were recorded. The results are in the book's **FishGills3** dataset.

   (a) Use StatKey to create a single histogram of *all* of the respiration rates, i.e., for all 3 groups combined. Describe the distribution, determine its summary statistics, and explain the relationship between its mean and median.

   The histogram is roughly symmetric, so the mean and median are fairly close. Summary statistics:

Mean	SD	Min	Q_1	Median	Q_3	Max
61.788	15.398	33	48	62.5	72	98

   (b) Using the histogram you just made, estimate the percentage of *all* of the respiration rates that are within 1 standard deviation of the overall average. Within 2 standard deviations? Within 3 standard deviations?

   Respiration rates between 46.38 and 77.176 are within 1 SD of the mean; there are about 56 rates in this interval, for a percentage of $\frac{56}{90} = .622 = 62.2\%$.

   Respiration rates between 30.982 and 92.574 are within 2 SDs of the mean; there are 88 rates in this interval, for a percentage of $\frac{88}{90} = .977 = 97.7\%$.

   All of the respiration rates are within 3 SDs of the mean.

   (c) Use StatKey to create boxplots of the respiration rates for the 3 experimental groups.

   See StatKey...

   (d) Determine the average, standard deviation, and five-number summary for each group's respiration rates.

   Here they are:

Group	Mean	SD	Min	Q_1	Median	Q_3	Max
Low	68.5	16.235	44	55	65	85	98
Medium	58.667	14.284	33	46	59.5	69	83
High	58.167	13.777	37	45	58.5	68	85

   (e) Based on this data, summarize the effect of calcium levels of water on fish respiration rates.

   Respiration rates are higher, on average, for the *low* group; respiration rates are about the same for the *medium* and *high* groups.

2. The drug finasteride is marketed as Propecia to help protect against male pattern baldness, and it may also protect against prostate cancer. A large sample of healthy men over age 55 were randomly assigned to receive either a daily finasteride pill or a placebo. The study lasted 7 years; the men had annual checkups and a biopsy at the end of the study. Prostate cancer was found in 804 of the 4,368 men taking finasteride and in 1,145 of the 4,692 men taking a placebo.

   (a) This study was double-blind. What does that mean?

   Neither the subjects nor researchers knew who received finasteride or the placebo.

   (b) Was this an experiment or an observational study?

   This was an experiment: the men were randomly assigned to treatment and control.

   (c) Summarize the results in a two-way table and include the row and column totals.

	Prostate cancer	No prostate cancer	Total
Finasteride	804	3564	4368
Placebo	1145	3547	4692
Total	1949	7111	9060

   (d) What percentage of the men in the study received finasteride? What percentage the men in the study had prostate cancer? Use correct notation for both proportions.

$$\text{Finasteride} : \hat{p} = \frac{4368}{9060} = 0.482 = 48.2\%$$

$$\text{Prostate cancer} : \hat{p} = \frac{1949}{9060} = 0.215 = 21.5\%$$

   (e) Compare the prostate cancer proportions for the two groups using appropriate notation. Does finasteride appear to offer some protection against prostate cancer?

$$\text{Finasteride} : \hat{p}_F = \frac{804}{4368} = 0.184 = 18.4\%$$

$$\text{Placebo} : \hat{p}_P = \frac{1145}{4692} = 0.244 = 24.4\%$$

   The prostate cancer rate was 6% lower for the finasteride group, so it does appear to offer some protection.

   (f) Is there evidence that finasteride protects against prostate cancer? State the relevant null and alternative hypotheses, use StatKey to create a randomization distribution based on this data and the null hypothesis, obtain the p-value, and state your conclusion clearly.

   The hypotheses are

$$H_0 : p_F = p_P \quad \text{versus} \quad H_a : p_F < p_P \ ,$$

   and the p-value based on the randomization distribution is practically 0. We have very strong evidence that finasteride protects against prostate cancer.

3. The **USStates** dataset includes the variables *HouseholdIncome*, the mean household income for a state (in thousands of dollars), and *College*, the percentage of a state's population age 25 or older who graduated from college. You are going to use this data to investigate the relationship between *College* and *HouseholdIncome*; read through all of the parts of this question to get your bearings before you do anything!

   (a) Use StatKey to create a scatterplot of *College* versus *HouseholdIncome*. Describe the trend shown and determine the correlation between these two variables.

   There is a positive association, and the correlation is $r = 0.686$.

   (b) What is the regression equation for predicting *College* from *HouseholdIncome*?

   $$\widehat{\text{College}} \;=\; 5.277 + 0.48 \times \text{HouseholdIncome}$$

   (c) What does the slope of this equation tell you? Interpret it in context.

   When the average household income increases by \$1000, the percentage of a state's population age 25 or older who graduated from college increases by 0.48 percent, on average.

   (d) What does the intercept of this equation tell you? Interpret it in context.

   If the average household income were zero – which makes no sense – the intercept says that the predicted percentage of college graduates would be 5.277%. The intercept makes no sense by itself in this problem.

   (e) (2 points) If the average household incomes for two states differ by \$10,000, how do their college graduation percentages differ, on average?

   $$10 \times 0.48 = 4.8\%$$

   (f) For a state with a mean household income of \$50,000, what is the predicted percentage of adults over 25 who have graduated from college?

   $$5.277 + 0.48 \times 50 \;=\; 29.277\%$$

4. The book's **Atlanta Commute (distance)** dataset (available directly on StatKey) provides the commute distances in miles for 500 randomly chosen people who work in the Atlanta metropolitan area.

   (a) Use StatKey to obtain the mean and standard deviation of the commute distances in this sample.

   $$\bar{x} = 18.156\,, \quad s = 13.798$$

   (b) Use StatKey to create 5000 bootstrap samples and a bootstrap distribution of sample means. What are the mean and standard error of your bootstrap distribution?

   The average for mine is 18.154, and the SE is 0.614 (results will vary a bit).

   (c) Use your bootstrap distribution to determine 95% and 99% confidence intervals for the average commute distance of workers in the Atlanta metropolitan area. Determine the margin of error for each of these confidence intervals. What happens to the margin of error as the confidence level increases?

   $$95\% : 16.971 \leq \mu \leq 19.376\,, \quad \text{MOE} \approx 1.2$$
   $$99\% : 16.620 \leq \mu \leq 19.815\,, \quad \text{MOE} \approx 1.6$$

   (d) Based on your bootstrap distribution and your interval estimates, which of the following is *most likely* and which is *least likely*?

      i. the average commute distance for people who work in the Atlanta metropolitan area is less than 20 miles

      Most likely! See the CI.

      ii. the average commute distance for people who work in the Atlanta metropolitan area is around 20 miles

      iii. the average commute distance for people who work in the Atlanta metropolitan area is greater than 20 miles

      Least likely! See the CI.

5. To analyze how well lie detectors perform when subjects are stressed, 48 randomly chosen subjects were connected to a lie detector and asked to read true statements out loud while receiving an electric shock. The lie detector incorrectly reported that 27 of the 48 participants were lying.

   Note: For this problem, you will have to enter the counts directly into StatKey. Do not use any of the book's available datasets to answer the following questions!

   (a) Use StatKey to determine the best estimate, based on this data, of the proportion of times the lie detector yields false positives, i.e., inaccurately reports deception.

   $$\hat{p} = 0.5625 = 56.25\%$$

   (b) Use StatKey to create 5000 bootstrap samples and a bootstrap distribution of sample proportions. What are the center and standard error of your bootstrap distribution?

   The center for mine is 0.562, and the SE is 0.072 (results will vary a bit).

   (c) Use StatKey and your bootstrap distribution to find a 95% confidence interval for the overall percentage of false positives reported by the lie detector. What is the margin of error?

   $$95\% \,:\, 0.417 = 41.7\% \;\le\; p \;\le\; 0.698 = 69.8\%\,, \quad \text{MOE} \approx .14 = 14\%$$

   (d) Does this sample provide evidence that lie detectors give inaccurate results more than half the time when subjects are stressed? State the relevant null and alternative hypotheses, use StatKey to create a randomization distribution based on this sample and H_0, obtain the p-value, and state your conclusion clearly.

   $$H_0 \,:\, p = 0.5\,, \quad H_a \,:\, p > 0.5$$

   Using $\hat{p} = .5625$ as the cutoff gives a p-value of around $0.25 = 25\%$; using $\hat{p} = .563$ as the cutoff gives a p-value of around $0.16 = 16\%$. In either case, we cannot reject the null hypothesis; as far as we know from this data, lie detectors only give inaccurate results half the time.