Chapter 1. Exercises

1. Use Table 1.3 to find the approximate quartiles of the distribution of the number of cigarettes smoked per day for the mothers in the CHDS who smoked during their pregnancy.

2. Combine the last four categories in Table 1.3 of the distribution of the number of cigarettes smoked by the smoking mothers in the CHDS. Make a new histogram using the collapsed table. How has the shape changed from the histogram in Figure 1.4? Explain.

3. Consider the histogram of father's age for the fathers in the CHDS (Figure 1.14). The bar over the interval from 35 to 39 years is missing. Find its height.

4. Consider the normal quantile plots of father's height and weight for fathers in the CHDS (Figure 1.15). Describe the shape of the distributions.

5. Following are the quantiles at 0.05, 0.10, . . ., 0.95 for the gestational ages of the babies in the CHDS. Plot these quantiles against those of the uniform distribution on $(0, 1)$. Describe the shape of the distribution of gestational age in comparison to the uniform.

   252, 262, 267, 270, 272, 274, 276, 277, 278, 280, 281, 283, 284, 286, 288, 290, 292, 296, 302.

6. Use the normal approximation to estimate the proportion of mothers in the CHDS between 62 and 64 inches tall to the nearest half inch (i.e., between 61.5 and 64.5 inches). The average height is 64 inches and the SD is 2.5 inches.

7. In the Missouri study, the average birth weight for babies born to smokers is 3180 grams and the SD 500 grams, and for nonsmokers the average is 3500 grams and the SD 500 grams. Consider a baby who is born to a smoker. If the baby's weight is 2 SDs below average weighs, then the baby weighs ________ grams. Suppose another baby weighs this same number of grams, but is born to a nonsmoker. This baby has a weight that falls ________ SDs below the average of its group. According to the normal approximation, approximately what percentage of babies born to nonsmokers are below this weight?

8. Suppose there are 100 observations from a standard normal distribution. What proportion of them would you expect to find outside the whiskers of a box-and-whisker plot?

9. Make a table for marital status that gives the percentage of smokers and nonsmokers in each marital category for the mothers in the Missouri study (Table 1.6).

10. Make a segmented bar graph showing the percentage at each education level for both smokers and nonsmokers for the mothers in the Missouri study (Table 1.6).

11. Make a bar graph of age and smoking status for the mothers in the Missouri study (Table 1.6). For each age group, the bar should denote the percentage of mothers in that group who smoke. How are age and smoking status related? Is age a potential confounding factor in the relationship between a mother's smoking status and her baby's birth weight?

12. In the Missouri study, the average birth weight for babies born to smokers is 3180 grams and the SD is 500 grams. What is the average and SD in ounces? There are 0.035 ounces in 1 gram.

13. Consider a list of numbers $x_1, \ldots, x_n$. Shift and rescale each x_i as follows:

    $$y_i \ = \ a + bx_i.$$

    Find the new average and SD of the list $y_1, \ldots y_n$ in terms of the average and SD of the original list $x_1, \ldots, x_n$.

14. Consider the data in Exercise 13. Express the median and IQR of $y_1, \ldots, y_n$ in terms of the median and IQR of $x_1, \ldots, x_n$. For simplicity, assume $y_1 < y_2 < \cdots < y_n$ and assume n is odd.

15. For a list of numbers $x_1, \ldots, x_n$ with $x_1 < x_2 \cdots < x_n$, show that by replacing x_n with another number, the average and SD of the list can be made arbitrarily large. Is the same true for the median and IQR? Explain.

16. Suppose there are n observations from a normal distribution. How could you use the IQR of the list to estimate σ?

17. Suppose the quantiles y_q of a $\mathcal{N}(\mu, \sigma^2)$ distribution are plotted against the quantiles z_q of a $\mathcal{N}(0, 1)$ distribution. Show that the slope and intercept of the line of points are σ and μ, respectively.

18. Suppose $X_1, \ldots, X_n$ form a sample from the standard normal. Show each of the following:

    (a) $\Phi(X_1), \ldots \Phi(X_n)$ is equivalent to a sample from a uniform distribution on $(0, 1)$. That is, show that for X a random variable with a standard normal distribution,

    $$P(\Phi(X) \leq q) = q.$$

    (b) Let $U_1, \ldots, U_n$ be a sample from a uniform distribution on $(0, 1)$. Explain why

    $$E(U_{(k)}) = \frac{k}{n+1} ,$$

    where $U_{(1)} \leq \ldots \leq U_{(n)}$ are the ordered sample.

    (c) Use (a) and (b) to explain why $X_{(k)} \approx z_{k/n+1}$.

19. Prove that $\bar{x}$ is the constant that minimizes the following squared error with respect to c:

    $$\sum_{i=1}^{n} (x_i - c)^2.$$

20. Prove that the median $\tilde{x}$ of $x_1, \ldots, x_n$ is the constant that minimizes the following absolute error with respect to c:

    $$\sum_{i=1}^{n} |x_i - c|.$$

    You may assume that there are an odd number of distinct observations. *Hint*: Show that if $c < c_o$, then

    $$\sum_{i=1}^{n} |x_i - c_o| = \sum_{i=1}^{n} |x_i - c| + (c - c_0)(r - s) + 2 \sum_{x \in (c, c_o)} (c - x_i) ,$$

    where $r =$ number of $x_i \geq c_o$, and $s = n - r$.