Problem set 02. Statistics and probability theory

Exercise 1. Mean versus median. A small accounting firm pays each of its five clerks \$35,000, two junior accountants \$80,000 each, and the firm's owner \$320,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary?

Exercise 2. How does the median change? The firm in the previous exercise gives no raises to the clerks and junior accountants, while the owner's take increases to \$455,000. How does this change affect the mean? How does it affect the median?

Exercise 3. Distributions for time spent studying. We asked the students in a large first-year college class how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 women and 30 men from the class:

Women					Men				
180	120	180	360	240	90	120	30	90	200
120	180	120	240	170	90	45	30	120	75
150	120	180	180	150	150	120	60	240	300
200	150	180	150	180	240	60	120	60	30
120	60	120	180	180	30	230	120	95	150
90	240	180	115	120	0	200	120	120	180

The most common methods for formal comparison of two groups use $\bar{x}$ and s to summarize the data. We wonder if this is appropriate here. Make the stemplot for this data set.
(a) What kinds of distributions are best summarized by $\bar{x}$ and s? It isn't easy to decide whether small data sets with irregular distributions fit the criteria.
(b) Each set of study times appears to contain a high outlier. Are these points flagged as suspicious by the 1.5 × IQR rule? How much does removing the outlier change $\bar{x}$ and s for each group? The presence of outliers makes us reluctant to use the mean and standard deviation for these data unless we remove the outliers on the grounds that these students were exaggerating.

Exercise 4. Hummingbirds and flowers. Different varieties of the tropical flower Heliconia are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the form of the hummingbirds' beaks have evolved to match each other. Here are data on the lengths in millimeters of three varieties of these flowers on the island of Dominica:

H. bihai							
47.12	46.75	46.81	47.12	46.67	47.43	46.44	46.64
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36

H. caribaea red							
41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57
39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07
38.10	37.97	38.79	38.23	38.87	37.78	38.01	

H. caribaea yellow							
36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.1
35.17	36.82	36.66	35.68	36.03	34.57	34.63	

Make boxplots to compare the three distributions. Report the five-number summaries along with your graph. What are the most important differences among the three varieties of flower?

Exercise 5. A uniform distribution. If you ask a computer to generate “random numbers” between 0 and 1, you will get observations from a uniform distribution. Figure 1.37 graphs the density curve for a uniform distribution. Use areas under this density curve to answer the following questions.

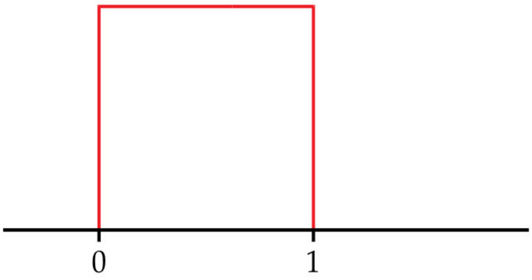

FIGURE 1.37 The density curve of a uniform distribution, for Exercise 1.108.

(a) Why is the total area under this curve equal to 1?
(b) What proportion of the observations lie below 0.35?
(c) What proportion of the observations lie between 0.35 and 0.65?

Exercise 6. Use a different range for the uniform distribution. Many random number generators allow users to specify the range of the random numbers to be produced. Suppose that you specify that the outcomes are to

be distributed uniformly between 0 and 4. Then the density curve of the outcomes has constant height between 0 and 4, and height 0 elsewhere.
(a) What is the height of the density curve between 0 and 4? Draw a graph of the density curve.
(b) Use your graph from (a) and the fact that areas under the curve are proportions of outcomes to find the proportion of outcomes that are less than 1.
(c) Find the proportion of outcomes that lie between 0.5 and 2.5.

Exercise 7. Find the mean, the median, and the quartiles. What are the mean and the median of the uniform distribution in Figure 1.37? What are the quartiles?

Exercise 8. Horse pregnancies are longer. Bigger animals tend to carry their young longer before birth. The length of horse pregnancies from conception to birth varies according to a roughly Normal distribution with mean 336 days and standard deviation 3 days. Use the 68-95-99.7 rule to answer the following questions.
(a) Almost all (99.7%) horse pregnancies fall in what range of lengths?
(b) What percent of horse pregnancies are longer than 339 days?

Exercise 9. Park space and population. Below are data on park and open space in several U.S. cities with high population density. In this table, population is reported in thousands of people, and park and open space is called open space, with units of acres.

City	Population	Open space
Baltimore	651	5,091
Boston	589	4,865
Chicago	2,896	11,645
Long Beach	462	2,887
Los Angeles	3,695	29,801
Miami	362	1,329
Minneapolis	383	5,694
New York	8,008	49,854
Oakland	399	3,712
Philadelphia	1,518	10,685
San Francisco	777	5,916
Washington, D.C.	572	7,504

(a) Make a bar graph for population. Describe what you see in the graph.
(b) Do the same for open space.
(c) For each city, divide the open space by population. This gives rates: acres

of open space per thousand residents.
(d) Make a bar graph of the rates.
(e) Redo the bar graph that you made in part
(d) by ordering the cities by their open space to population rate.
(f) Which of the two bar graphs in (d) and (e) do you prefer? Give reasons for your answer.

Exercise 10. Compare two Normal curves. We use the fact that the distribution of scores for the 76,531 students who took the exam was approximately N(572, 51). These students were classified in a variety of ways, and summary statistics were reported for these different subgroups. When classified by gender, the scores for the women are approximately N(579, 49), and the scores for the men are approximately N(565, 55). Figure 1.43 gives the Normal density curves for these two distributions. Here is a possible description of these data: women score about 14 points higher than men on the ISTEP English/language arts exam. Critically evaluate this statement and then write your own summary based on the distributions displayed in Figure 1.43.

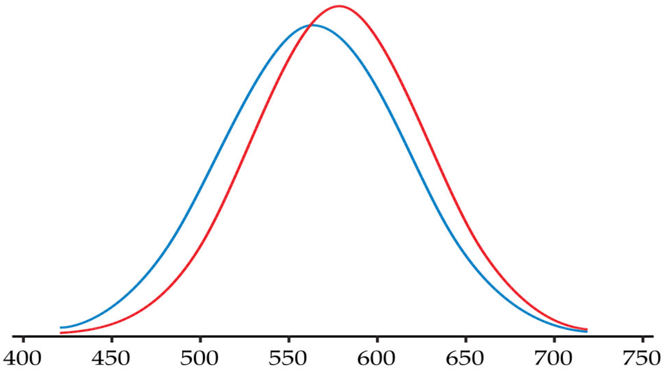

FIGURE 1.43 Normal density curves for ISTEP scores of women and men, for Example 1.53.

FIGURE 1.43 Normal density curves for ISTEP scores of women and men, for Example 1.53.

Exercise 11. Biological clocks. Many plants and animals have "biological clocks" that coordinate activities with the time of day. When researchers looked at the length of the biological cycle in the plant Arabidopsis by measuring leaf movements, they found that the length of the cycle is not always 24 hours. Further study discovered that cycle length changes systematically with north-south location.

Biological clock cycle lengths for a plant species in different locations

23.89	23.72	23.74	24.35	25.05	24.56	23.69	22.33	23.79	22.12
25.39	23.08	25.64	23.98	25.84	25.46	24.37	24.13	24.40	24.74
24.44	24.82	23.56	24.96	24.21	23.85	24.57	23.44	23.64	24.23
24.01	24.58	25.57	23.73	24.11	23.21	25.08	24.03	24.62	23.51
23.21	23.41	23.69	22.97	24.65	24.65	24.29	23.89	25.08	23.89
24.95	23.09	23.21	24.66	23.88	25.33	24.38	24.68	25.34	25.22
23.45	23.39	25.43	23.16	23.95	23.25	24.72	24.89	24.88	24.71
23.58	25.98	24.28	24.25	23.16	24.19	27.22	23.77	26.21	24.33
24.34	24.89	24.32	24.14	24.00	23.48	25.81	24.99	24.18	22.73
24.18	23.95	24.48	23.89	24.24	24.96	24.58	24.29	24.31	23.64
23.87	23.68	24.87	23.00	23.48	24.26	23.34	25.11	24.69	24.97
24.64	24.49	23.61	24.07	26.60	24.91	24.76	25.09	26.56	25.13
24.81	25.63	25.63	24.69	24.41	23.79	22.88	22.00	23.33	25.12
24.00	24.31	23.03	24.51	28.55	22.96	23.61	24.72	24.04	25.18
24.30	24.22	24.39	24.73	24.68	24.14	24.57	24.42	25.62	

Table 1.11 contains cycle lengths for 149 locations around the world. Describe the distribution of cycle lengths with a histogram and numerical summaries. In particular, how much variation is there among locations?

Exercise 12. Product preference. Product preference depends in part on the age, income, and gender of the consumer. A market researcher selects a large sample of potential car buyers. For each consumer, she records gender, age, household income, and automobile preference. Which of these variables are categorical and which are quantitative?

Exercise 13. Distance-learning courses. The 222 students enrolled in distance-learning courses offered by a college ranged from 18 to 64 years of age. The mode of their ages was 19. The median age was 31. Explain how this can happen.

Exercise 14. Norms for reading scores. Raw scores on behavioral tests are often transformed for easier comparison. A test of reading ability has mean 75 and standard deviation 10 when given to third-graders. Sixth-graders have mean score 82 and standard deviation 11 on the same test. To provide separate "norms" for each grade, we want scores in each grade to have mean 100 and standard deviation 20.
(a) What linear transformation will change third- grade scores x into new scores $x_{new} = a + bx$ that have the desired mean and standard deviation? (Use $b > 0$ to preserve the order of the scores.)

(b) Do the same for the sixth-grade scores.
(c) David is a third-grade student who scores 78 on the test. Find David's transformed score. Nancy is a sixth-grade student who scores 78. What is her transformed score? Who scores higher within his or her grade?
(d) Suppose that the distribution of scores in each grade is Normal. Then both sets of transformed scores have the $N(100, 20)$ distribution. What percent of third-graders have scores less than 78? What percent of sixth-graders have scores less than 78?

Exercise 15. Damage caused by tornados. The average damage caused by tornadoes in the states

State	Damage ($ millions)	State	Damage ($ millions)	State	Damage ($ millions)
Alabama	51.88	Louisiana	27.75	Ohio	44.36
Alaska	0.00	Maine	0.53	Oklahoma	81.94
Arizona	3.47	Maryland	2.33	Oregon	5.52
Arkansas	40.96	Massachusetts	4.42	Pennsylvania	17.11
California	3.68	Michigan	29.88	Puerto Rico	0.05
Colorado	4.62	Minnesota	84.84	Rhode Island	0.09
Connecticut	2.26	Mississippi	43.62	South Carolina	17.19
Delaware	0.27	Missouri	68.93	South Dakota	10.64
Florida	37.32	Montana	2.27	Tennessee	23.47
Georgia	51.68	Nebraska	30.26	Texas	88.60
Hawaii	0.34	Nevada	0.10	Utah	3.57
Idaho	0.26	New Hampshire	0.66	Vermont	0.24
Illinois	62.94	New Jersey	2.94	Virginia	7.42
Indiana	53.13	New Mexico	1.49	Washington	2.37
Iowa	49.51	New York	15.73	West Virginia	2.14
Kansas	49.28	North Carolina	14.90	Wisconsin	31.33
Kentucky	24.84	North Dakota	14.69	Wyoming	1.78

and the estimated amount of oil recovered from different oil wells

21.7	53.2	46.4	42.7	50.4	97.7	103.1	51.9
43.4	69.5	156.5	34.6	37.9	12.9	2.5	31.4
79.5	26.9	18.5	14.7	32.9	196.0	24.9	118.2
82.2	35.1	47.6	54.2	63.1	69.8	57.4	65.6
56.4	49.4	44.9	34.6	92.2	37.0	58.8	21.3
36.6	64.9	14.8	17.6	29.1	61.4	38.6	32.5
12.0	28.3	204.9	44.5	10.3	37.7	33.7	81.1
12.1	20.1	30.5	7.1	10.1	18.0	3.0	2.0

Make histograms of those distributions and show that they both are right-skewed.