DESCRIBING DATA WITH NUMBERS

- A numerical summary of a distribution should report at least its **center** and its **spread** or **variability.**
- The **mean** $\bar{x}$ and the **median** M describe the center of a distribution in different ways. The mean is the arithmetic average of the observations, and the median is the midpoint of the values.
- When you use the median to indicate the center of the distribution, describe its spread by giving the **quartiles.** The **first quartile** Q_1 has one-fourth of the observations below it, and the **third quartile** Q_3 has three-fourths of the observations below it.
- The **five-number summary** consisting of the median, the quartiles, and the smallest and largest individual observations provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.
- **Box plots** based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data.
- The **variance** s^2 and especially its square root, the **standard deviation** s**,** are common measures of spread about the mean as center. The standard deviation s is zero when there is no spread and gets larger as the spread increases.
- A **resistant measure** of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are. The median and quartiles are resistant, but the mean and the standard deviation are not.
- The mean and standard deviation are good descriptions for symmetric distributions without outliers. They are most useful for the Normal distributions introduced in the next chapter. The five-number summary is a better description for skewed distributions.
- Numerical summaries do not fully describe the shape of a distribution. Always plot your data.
- A statistical problem has a real-world setting. You can organize many problems using the four steps **state, plan, solve,** and **conclude.**

EXERCISES

Weight of newborns. Here is the distribution of the weight at birth for all babies born in Kenya in 2011:

Weight (grams)	Count
Less than 500	6,599
500 to 999	23,864
1,000 to 1,499	31,325
1,500 to 1,999	66,453
2,000 to 2,499	210,324
2,500 to 2,999	748,042
3,000 to 3,499	1,596,944
3,500 to 3,999	1,114,887
4,000 to 4,499	289,098
4,500 to 4,999	42,119
5,000 to 5,499	4,715

(a) For comparison with other years and with other countries, we prefer a histogram of the percents in each weight class rather than the counts. Explain why. Construct a distribution table showing the % of babies in each class.

(b) How many babies were there? Make a histogram of the distribution, using percents on the vertical scale.
(c) What are the positions of the median and quartiles in the ordered list of all birth weights? In which weight classes do the median and quartiles fall?

Guinea pig survival times. Here are the survival times in days of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment. Survival times, whether of machines under stress or cancer patients after treatment, usually have distributions that are skewed to the right.

43	45	53	56	56	57	58	66	67	73	74	79
80	80	81	81	81	82	83	83	84	88	89	91
91	92	92	97	99	99	100	100	101	102	102	102
103	104	107	108	109	113	114	118	121	123	126	128
137	138	139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511	522	598

(a) Graph the distribution and describe its main features. Does it show the expected right skew?
(b) Which numerical summary would you choose for these data? Calculate your chosen summary. How does it reflect the skew-ness of the distribution?
(c) Compute the standard deviation of the graph and also locate the median, 1^{st} quartile and third quartile and hence get the inter quartile range. Make a box plot for the data.