Chapter 12

Describing Distributions with Numbers

Lecture Slides

Case Study: Describing Distributions with Numbers 1

Does education pay? We are told that people with more education earn more, on the average, than people with less education.

How much more? How can we answer this question? Data on income can be found at the Census Bureau website.

The data are estimates for the year 2013 of the total incomes of 136,641,000 people aged 25 and over, with earnings, and are based on the results of the Current Population Survey in 2014.

Case Study: Describing Distributions with Numbers 2

The website gives the income distribution for each of several education categories.

It gives the number of people in each of several education categories who earned between $1 and $2499, between $2500 and $4999, up to between $97,500 and $99,999, and $100,000 and over. That is a lot of information.

A histogram could be used to display the data, but are there simple ways to summarize the information with just a few numbers that allow us to make sensible comparisons?

Case Study: Describing Distributions with Numbers 3

In this chapter, we will learn several ways to summarize large data sets with a few numbers. By the end of this chapter, with these new methods for summarizing large data sets, you will be able to provide an answer to whether education really pays.

Describing Distributions with Numbers 1

Baseball has a rich tradition of using statistics to summarize and characterize the performance of players. We begin by investigating ways to summarize the performance of the greatest home-run hitters of all time. In the summer of 2007, Barry Bonds shattered the career home-run record, breaking the previous record set by Hank Aaron. Here are his home-run counts for the years 1986 (his rookie year) to 2007 (his final season):

1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
16	25	24	19	33	25	34	46	37	33	42
1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
40	37	34	49	73	46	45	45	5	26	28

Describing Distributions with Numbers 2

The shape of the distribution is a bit irregular, but we see that it has one high outlier, and if we ignore this outlier, we might describe it as slightly skewed to the left with a single peak.

The outlier is Bonds's record season in 2001.

stem	leaf
0	5
1	69
2	45568
3	334477
4	0255669
5	
6	
7	3

Key: A stem of 1 and a leaf of 6 means 16 home runs.

Moore/Notz, *Statistics: Concepts and Controversies*, 10e,

Median and Quartiles 1

A simple and effective way to describe center and variability is to give the median and the quartiles.

The median is the midpoint, the value that separates the smaller half of the observations from the larger half.

The quartiles get their name because, with the median, they divide the observations into quarters: one-quarter of the observations lie below the first quartile, one half lies below the median, and three-quarters lie below the third quartile. That's the idea. To actually get numbers, we need a rule that makes the idea exact.

Example: Finding the median

We might compare Bonds's career with that of Hank Aaron, the previous holder of the career record. Here are Aaron's home-run counts for his 23 years in baseball:

13 27 26 44 30 39 40 34 45 44 24 32 44 39 29 44 38 47 34 40 20 12 10

To find the median, first arrange them in order from smallest to largest:

10 12 13 20 24 26 27 29 30 32 34 **34** 38 39 39 40 40 44 44 44 44 45 47

There is an odd number of observations, so the median is the middle value.

Example: Finding the median (continued)

How does this compare with Bonds's record? Here are Bonds's 22 home run counts, arranged in order from smallest to largest:

n is even, so there is no one middle observation. There is a middle pair.

Take the median to be halfway between this middle pair.

Median is M = $\frac{34+34}{2} = 34$

Median and Quartiles 2

The **median (*M*)** is the midpoint of a distribution, the number such that one half of the observations is smaller and the other half is larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations *n* is odd, the median (*M)* is the center observation in the ordered list. Find the location of the median by counting $(n + 1)/2$ observations up from the bottom of the list.
3. If the number of observations *n* is even, the median (*M*) is the average of the two center observations in the ordered list. The location of the median is again $(n + 1)/2$ from the bottom of the list.

Median and Quartiles 3

The Census Bureau website provides data on income inequality. For example, it tells us that in 2013 the median income of Hispanic households was $40,963. That's helpful but incomplete.

Do most Hispanic households earn close to this amount, or are the incomes very variable?

The simplest useful description of a distribution consists of both a measure of center and a measure of variability.

Median and Quartiles 4

If we choose the median (the midpoint) to describe center, the quartiles (in particular, the difference between the quartiles) provide natural descriptions of variability.

The idea is clear: find the points one-quarter and three-quarters up the ordered list of observations. Again, we need a rule to make the idea precise. The rule for calculating the quartiles uses the rule for the median.

Median and Quartiles 5

To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median (M) in the ordered list of observations.
2. The first quartile (Q_1) is the median of the observations whose position in the ordered list is to the left of the location of the overall median. The overall median is not included in the observations considered to be to the left of the overall median.
3. The third quartile (Q_3) is the median of the observations whose position in the ordered list is to the right of the location of the overall median. The overall median is not included in the observations considered to be to the right of the overall median.

Example: Finding the quartiles

Find the quartiles for Hank Aaron's home run distribution.

10 12 13 20 24 26 27 29 30 32 34 **34** 38 39 39 40 40 44 44 44 44 45 47

There is an odd number of observations, so the median is the middle value.

The lower half of the data is:

10 12 13 20 24 **26** 27 29 30 32 34

The upper half of the data is:

38 39 39 40 40 **44** 44 44 44 45 47

Five-Number Summary and Boxplots 1

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

Minimum Q_1 M Q_3 Maximum

Five-Number Summary and Boxplots 2

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles.
- A line in the box marks the median.
- Lines extend from the box out to the smallest and largest observations.

Five-Number Summary and Boxplots 3

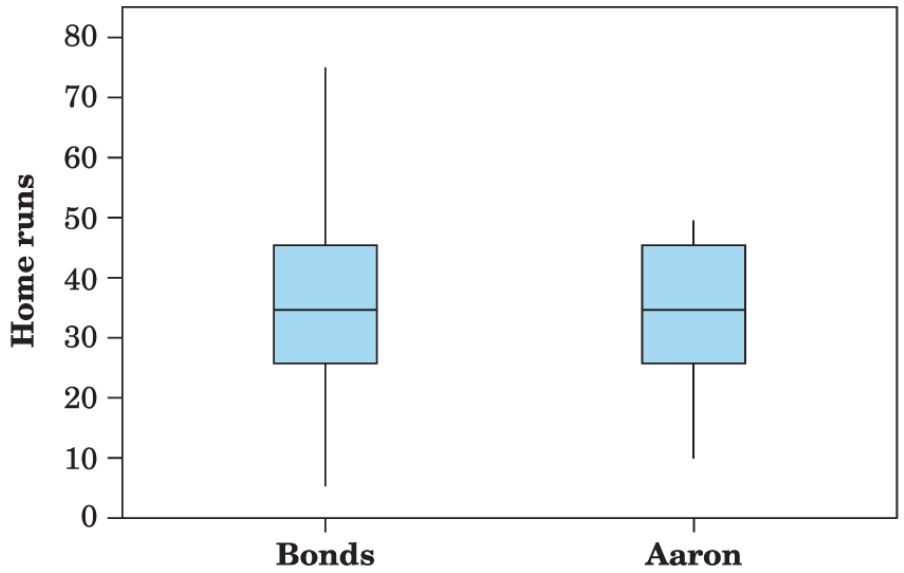
Boxplots can be drawn horizontally or vertically. Be sure to include a numerical scale in the graph.

When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the variability.

The quartiles (more precisely, the difference between the two quartiles) show the variability of the middle half of the data, and the extremes (the smallest and largest observations) indicate the variability of the entire data set.

Five-Number Summary and Boxplots 4

We see from Figure 12.2 that Bonds's usual performance, as indicated by the median and the box that marks the middle half of the distribution, is similar to that of Aaron. We also see that the distribution for Aaron is less variable than the distribution for Bonds.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Mean and Standard Deviation 1

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the standard deviation to measure variability.

The mean is the ordinary average of the observations.

The standard deviation gives the average distance of observations from the mean.

For now, just think of the standard deviation as “average distance from the mean” and leave the details to your calculator.

Mean and Standard Deviation 2

$$mean = \bar{x} = \frac{sum\ of\ observations}{n}$$

To find the standard deviation, *s*, of *n* observations:

1. Find the distance of each observation from the mean and square each of these distances.
2. Average the squared distances by dividing their sum by n − 1. This average squared distance is called the variance.
3. The standard deviation *s* is the square root of this average squared distance.

Example: Finding the mean and standard deviation 1

The numbers of home runs Barry Bonds hit in his 22 major league seasons are

16 25 24 19 33 25 34 46 37 33 42 40 37 34 49 73 46 45 45 5 26 28

To find the mean of these observations

$$\bar{x} = \frac{sum\ of\ observations}{n}$$

$$= \frac{16 + 25 + \cdots + 28}{22}$$

$$= 34.6\ home\ runs$$

Example: Finding the mean and standard deviation 2

To find the standard deviation of these observations:

Observation	Squared distance from mean
16	$(16 - 34.6)^2 = (-18.6)^2 = 345.96$
25	$(25 - 34.6)^2 = (-9.6)^2 = 92.16$
$\vdots$	
28	$(28 - 34.6)^2 = (-6.6)^2 = 43.56$
	sum $= 4139.12$

Example: Finding the mean and standard deviation 3

The average is $\frac{4139.12}{21} = 197.1$

Notice that we "average" by dividing by one less than the number of observations. Finally, the standard deviation is the square root of this number:

$$s = \sqrt{197.1} = 14.04 \; home \; runs$$

Mean and Standard Deviation 3

Properties of the standard deviation *s*

- *s* measures variability about the mean $\bar{x}$.
- Use *s* to describe the variability of a distribution only when you use $\bar{x}$ to describe the center.
- $s = 0$ only when there is no variability. This happens only when all observations have the same value. So standard deviation zero means no variability at all. Otherwise $s > 0$. As the observations become more variable about their mean, *s* gets larger.

Choosing Numerical Descriptions 1

mean vs. median

- The mean is strongly influenced by a few extreme observations.
- The median is not.

If the distribution is

- Symmetrical → mean = median
- Skewed right → mean > median
- Skewed left → mean < median

Choosing Numerical Descriptions 2

Choosing a summary:

The mean and standard deviation are strongly affected by outliers or by the long tail of a skewed distribution.

The median and quartiles are less affected.

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with outliers. Use $\bar{x}$ and s only for reasonably symmetric distributions that are free of outliers.

Choosing Numerical Descriptions 3

Numerical summaries do not disclose the presence of multiple peaks or gaps.

A picture will help you detect skewness and outliers.

Always start with a graph of your data

Statistics in Summary 1

- If we have data on a single quantitative variable, we start with a histogram or stemplot to display the distribution. Then we add numbers to describe the **center and variability** of the distribution.

- There are two common descriptions of center and variability: the **five-number summary** and the **mean** and **standard deviation**.

Statistics in Summary 2

The five-number summary consists of the **median** M, the midpoint of the observations, to measure center and the difference between the two **quartiles** Q1 and Q3 and the difference between the smallest and largest observations to describe variability.

A **boxplot** is a graph of the five-number summary.

The **mean**, $\bar{x}$, is the average of the observations.

Statistics in Summary 3

The **standard deviation**, *s*, measures variability as a kind of average distance from the mean, so use it only with the mean. The variance is the square of the standard deviation.

The mean and standard deviation can be changed a lot by a few outliers. The mean and median are the same for symmetrical distributions, but the mean moves farther toward the long tail of a skewed distribution.

In general, use the five-number summary to describe most distributions and the mean and standard deviation only for roughly symmetrical distributions.