Chapter 6 - The Standard Deviation as a Ruler and the Normal Model

Obj - SWBAT know what z-scores mean and be able to explain how extraordinary a standardized value may be by using a Normal model.

The Standard Deviation as a Ruler

- The trick in comparing very different-looking values is to use standard deviations as our rulers.
- The standard deviation tells us how the whole collection of values varies, so it's a natural ruler for comparing an individual to a group.
- As the most common measure of variation, the standard deviation plays a crucial role in how we look at data.

Standardizing with *z*-scores

- We compare individual data values to their mean, relative to their standard deviation using the following formula:

$$z = \frac{(y - \bar{y})}{s}$$

- We call the resulting values standardized values, denoted as z. They can also be called z-scores.

Standardizing with *z*-scores (cont.)

- Standardized values have no units.
- z-scores measure the distance of each data value from the mean in standard deviations.
- A negative z-score tells us that the data value is below the mean, while a positive z-score tells us that the data value is above the mean.

Benefits of Standardizing

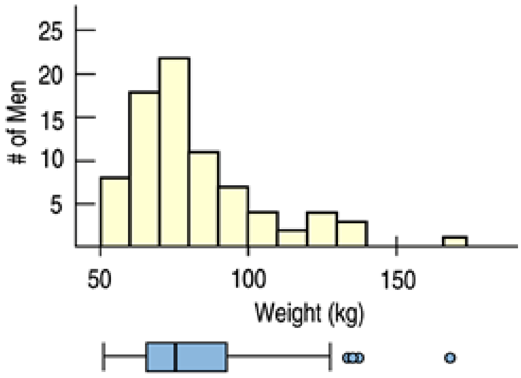
- Standardized values have been converted from their original units to the standard statistical unit of standard deviations from the mean.
- Thus, we can compare values that are measured on different scales, with different units, or from different populations.

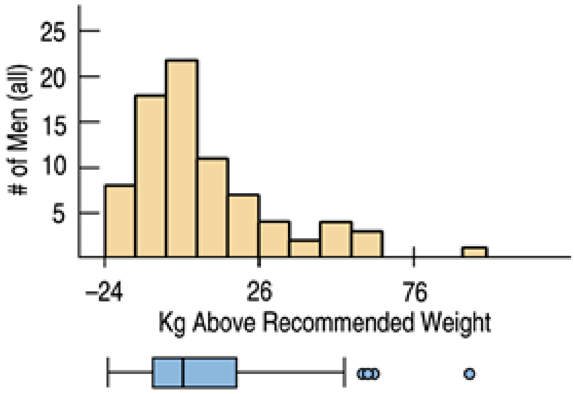
Benefits of Standardizing (cont'd)

- Shifting data:
- Adding (or subtracting) a constant amount to each value just adds (or subtracts) the same constant to (from) the mean.

Shifting Data (cont.)

- The following histograms show a shift from men's actual weights to kilograms above recommended weight:

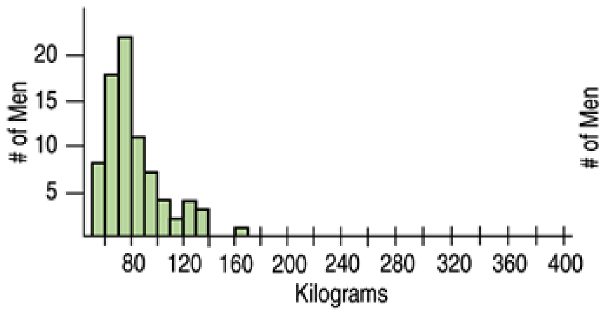


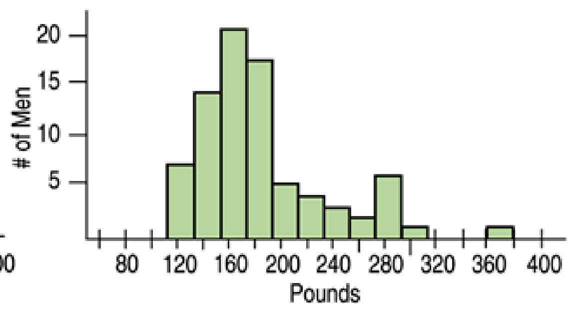


Rescaling Data

Rescaling Data (cont.)

- The men's weight data set measured weights in kilograms. If we want to think about these weights in pounds, we would rescale the data:





Back to *z*-scores

- Standardizing data into z-scores shifts the data by subtracting the mean and rescales the values by dividing by their standard deviation.
- Standardizing into z-scores does not change the shape of the distribution.

When Is a *z*-score **BIG**?

- A z-score gives us an indication of how unusual a value is because it tells us how far it is from the mean.

When Is a *z*-score **BIG**?

- How far from 0 does a z-score have to be to be interesting or unusual?
- There is no universal standard, but the larger a z-score is (negative or positive), the more unusual it is.
- Remember that a negative z-score tells us that the data value is below the mean, while a positive z-score tells us that the data value is above the mean.

When Is a *z*-score Big? (cont.)

- There is no universal standard for z-scores, but there is a model that shows up over and over in Statistics.
- This model is called the Normal model (You may have heard of “bell-shaped curves.”).

Distribution	Model

When Is a *z*-score Big? (cont.)

- There is a Normal model for every possible combination of mean and standard deviation.
- We write N(μ,σ) to represent a Normal model with a mean of μ and a standard deviation of σ.
- We use Greek letters because this mean and standard deviation are not numerical summaries of the data. They are part of the model. They don't come from the data. They are numbers that we choose to help specify the model.
- Such numbers are called parameters of the model.

When Is a *z*-score Big? (cont.)

- Summaries of data, like the sample mean and standard deviation, are written with Latin letters. Such summaries of data are called statistics.
- When we standardize Normal data, we still call the standardized value a z-score, and we write

$$z = \frac{y - \mu}{\sigma}$$

When Is a *z*-score Big? (cont.)

- Once we have standardized, we need only one model:
- The N(0,1) model is called the standard Normal model (or the standard Normal distribution).
- Be careful—don't use a Normal model for just any data set, since standardizing does not change the shape of the distribution.

- When we use the Normal model, we are assuming the distribution is Normal.
- We cannot check this assumption in practice, so we check the following condition:

Nearly Normal Condition: The shape of the data's distribution is unimodal and symmetric.

- This condition can be checked with a histogram or a Normal probability plot (to be explained later).

The 68-95-99.7 Rule

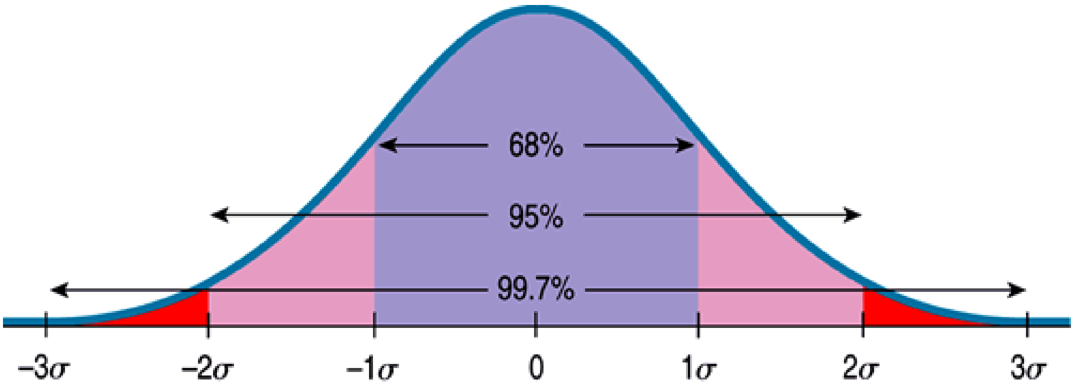
- Normal models give us an idea of how extreme a value is by telling us how likely it is to find one that far from the mean.
- We can find these numbers precisely, but until then we will use a simple rule that tells us a lot about the Normal model…

The 68-95-99.7 Rule (cont.)

- It turns out that in a Normal model:
- about 68%
- about 95%
- about 99.7% (almost all!)

The 68-95-99.7 Rule (cont.)

The following shows what the 68-95-99.7 Rule tells us:



The First Three Rules for Working with Normal Models

- Make a picture.
- Make a picture.
- Make a picture.
- And, when we have data, make a histogram to check the Nearly Normal Condition to make sure we can use the Normal model to model the distribution.

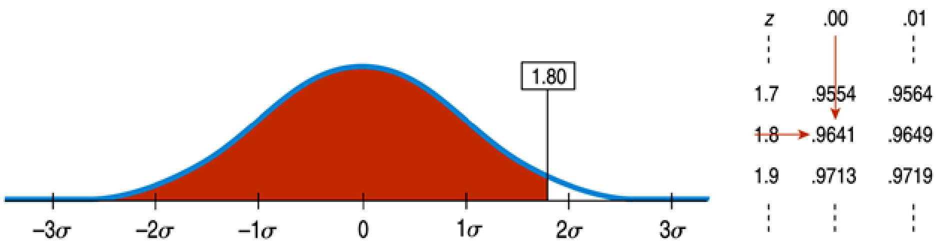
Finding Normal Percentiles by Hand

- When a data value doesn't fall exactly 1, 2, or 3 standard deviations from the mean, we can look it up in a table of Normal percentiles.
- Table Z in Appendix D provides us with normal percentiles, but many calculators and statistics computer packages provide these as well.

Finding Normal Percentiles by Hand (cont.)

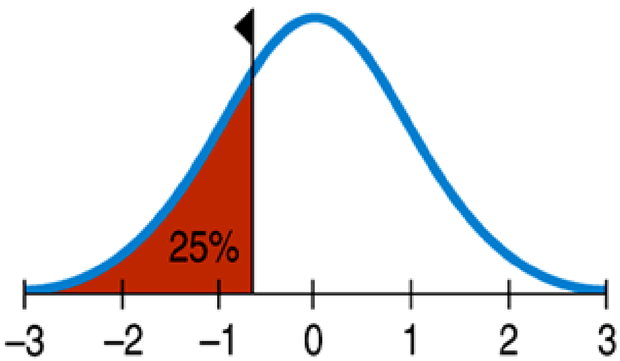
Table Z is the standard Normal table. We have to convert our data to z-scores before using the table.

The figure shows us how to find the area to the left when we have a z-score of 1.80:



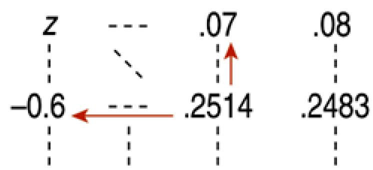
From Percentiles to Scores: *z* in Reverse

- Sometimes we start with areas and need to find the corresponding z-score or even the original data value.
- Example: What z-score represents the first quartile in a Normal model?



From Percentiles to Scores: *z* in Reverse (cont.)

- Look in Table Z for an area of 0.2500.
- The exact area is not there, but 0.2514 is pretty close.



- This figure is associated with z = –0.67, so the first quartile is 0.67 standard deviations below the mean.

Are You Normal? How Can You Tell?

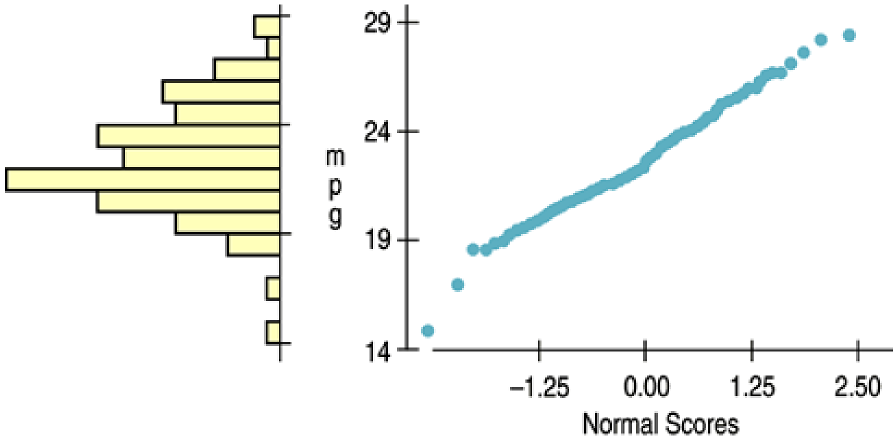
- When you actually have your own data, you must check to see whether a Normal model is reasonable.
- Looking at a histogram of the data is a good way to check that the underlying distribution is roughly unimodal and symmetric.

Are You Normal? How Can You Tell? (cont.)

- A more specialized graphical display that can help you decide whether a Normal model is appropriate is the Normal probability plot.
- If the distribution of the data is roughly Normal, the Normal probability plot approximates a diagonal straight line. Deviations from a straight line indicate that the distribution is not Normal.

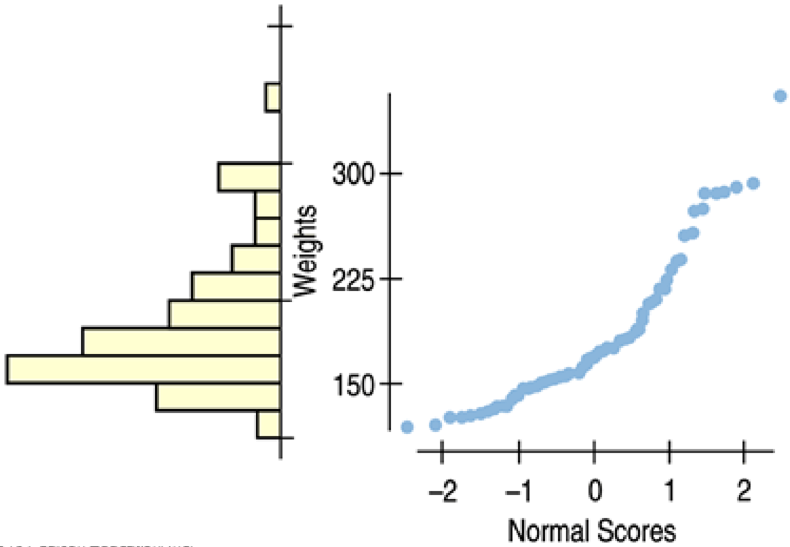
Are You Normal? How Can You Tell? (cont.)

Nearly Normal data have a histogram and a Normal probability plot that look somewhat like this example:



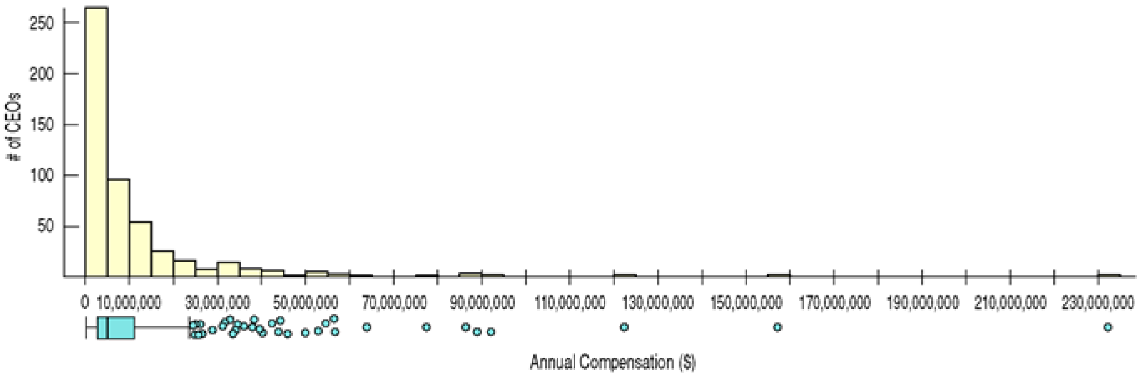
Are You Normal? How Can You Tell? (cont.)

A skewed distribution might have a histogram and Normal probability plot like this:



What Can Go Wrong?

- Don't use a Normal model when the distribution is not unimodal and symmetric.



What Can Go Wrong? (cont.)

- Don't use the mean and standard deviation when outliers are present—the mean and standard deviation can both be distorted by outliers.
- Don't round off too soon.
- Don't round your results in the middle of a calculation.
- Don't worry about minor differences in results.

What have we learned?

- The story data can tell may be easier to understand after shifting or rescaling the data.
- Shifting data by adding or subtracting the same amount from each value affects measures of center and position but not measures of spread.
- Rescaling data by multiplying or dividing every value by a constant changes all the summary statistics—center, position, and spread.

What have we learned? (cont.)

- *We've learned the power of standardizing data.*
- *Standardizing uses the SD as a ruler to measure distance from the mean (z-scores).*
- *With z-scores, we can compare values from different distributions or values based on different units.*
- *z-scores can identify unusual or surprising values among data.*

What have we learned? (cont.)

- We see the importance of Thinking about whether a method will work:
- Normality Assumption: We sometimes work with Normal tables (Table Z). These tables are based on the Normal model.
- Data can't be exactly Normal, so we check the Nearly Normal Condition by making a histogram (is it unimodal, symmetric and free of outliers?) or a normal probability plot (is it straight enough?).

Suppose the class took a 40-point quiz. Results show a mean score of 30, median 32, IQR 8, SD 6, min 12, and Q_1 27. (Suppose YOU got a 35.) What happens to each of the statistics if…

•I decide to weight the quiz as 50 points, and will add 10 points to every score.

•I decide to weight the quiz as 80 points, and double each score.

•I decide to count the quiz as 100 points; I'll double each score and add 20 points.

Statistic	original (y)			
mean	30			
median	32			
IQR	8			
SD	6			
minimum	12			
Q_1	27			
your score	35			

Let's talk about scoring the decathlon. Silly example, but suppose two competitors tie in each of the first eight events. In the ninth event, the high jump, one clears the bar 1 in. higher. Then in the 1500-meter run the other one runs 5 seconds faster. Who wins? It boils down to knowing whether it is harder to jump an inch higher or run 5 seconds faster. So consider the three athletes' performances shown below in a three event competition. Note that each placed first, second, and third in an event. Who gets the gold medal? Who turned in the most remarkable performance of the competition?

Events

Competitor	100 m Dash	Shot Put	Long Jump
A	10.1 sec	66'	26'
B	9.9 sec	60'	27'
C	10.3 sec	63'	27'3"
mean	10 sec	60'	26'
St Dev	0.2 sec	3'	6"

Competitor	100 m Dash	Shot Put	Long Jump
A	**10.1 sec**	**66'**	**26'**
B	**9.9 sec**	**60'**	**27'**
C	**10.3 sec**	**63'**	**27'3"**

Sketch the following Normal Models using the 69-95-99.7 Rule:

- birth weights of babies N(7.6lb, 1.3lb)

- ACT scores at a certain college N(21.2, 4.4)

Suppose a Normal model describes the fuel efficiency of cars currently registered in your state. The mean is 24 mpg, with a standard deviation of 6 mpg.

- Sketch and describe the Normal model.

- What percent of cars get less than 15 mpg?

- What percent of all cars get 20 to 30 mpg?

- What percent of cars get more than 40 mpg?

- Describe the fuel efficiency of the worst 20% of all cars.

- What gas mileage represents the third quartile?

- Describe the gas mileage of the most efficient 5% of all cars.

- An ecology group is lobbying for a national goal calling for no more than 10% of all cars to be under 20 mpg. If the standard deviation does not change what average fuel efficiency must be attained?

- Car manufacturers argue that they cannot raise the average that much – they believe they can only get to 26 mpg. What standard deviation would allow them to meet the "only 10% under 20 mpg" goal?

- What change in the fuel economy of cars would achieving that standard deviation bring about? What are the advantages and disadvantages?