Starter Ch. 6: A z-score Analysis

Your Statistics teacher has announced that the lower of your two tests will be dropped. You got a 90 on test 1 and an 85 on test 2. You're all set to drop the 85 until he announces that he grades "*on a curve*." He standardized the scores in order to decide which is the lower one. If the mean on the first test was 88 with a standard deviation of 4 and the mean on the second was 75 with a standard deviation of 5,

a) Which one will be dropped?

b) Does this seem "fair"?

1

Starter Ch. 6 2001 #1ab

The summary statistics for the number of inches of rainfall in Los Angeles for 117 years, beginning in 1877, are shown below.

N	MEAN	MEDIAN	TRMEAN	STDEV	SE MEAN
117	14.941	13.070	14.416	6.747	0.624

MIN	MAX	Q1	Q3
4.850	38.180	9.680	19.250

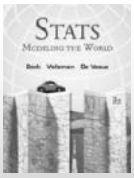
(a) Describe a procedure that uses these summary statistics to determine whether there are outliers.

(b) Are there outliers in these data? _____

Justify your answer based on the procedure that you described in part (a).

Chapter 6

The Standard Deviation as a Ruler and the 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.

7

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.

8

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.

9

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.

10

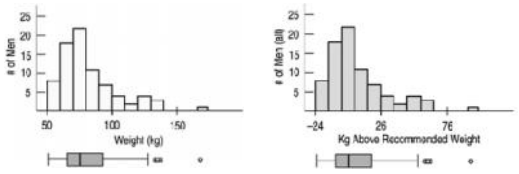
Shifting Data

- Shifting data:
  - Adding (or subtracting) a *constant* amount to each value just adds (or subtracts) the same constant to (from) the mean. This is true for the median and other measures of position too.
  - In general, ***adding a constant to every data value adds the same constant to measures of center and percentiles, but leaves measures of spread unchanged***.

11

Shifting Data (cont.)

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



12

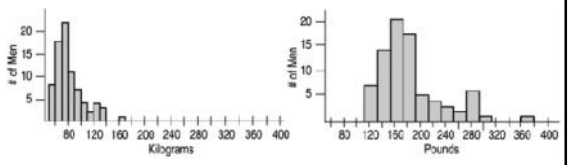
Rescaling Data

- Rescaling data:
  - When we **divide or multiply** all the data values by any constant value, **all measures of position** (such as the mean, median and percentiles) and **measures of spread** (such as the range, IQR, and standard deviation) are **divided and multiplied by that same constant value**.

13

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:



14

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.
  - Standardizing into *z*-scores changes the **center** by making the **mean 0**.
  - Standardizing into *z*-scores changes the **spread** by making the **standard deviation 1**.

15

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.
- A data value that sits right at the mean, has a *z*-score equal to 0.
- A *z*-score of 1 means the data value is 1 standard deviation above the mean.
- A *z*-score of –1 means the data value is 1 standard deviation below the mean.

16

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.

17

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.").
- Normal models are appropriate for distributions whose shapes are unimodal and roughly symmetric.
- These distributions provide a measure of how extreme a *z*-score is.

18

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.

19

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

$$z = \frac{y - «}{±}$$

20

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.

21

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

- 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).

22

The 68-95-99.7 Rule

The 68–95–99.7 Rule

In the Normal distribution with mean μ and standard deviation σ:

- Approximately 68% of the observations fall within σ of the mean μ.
- Approximately 95% of the observations fall within 2σ of μ.
- Approximately 99.7% of the observations fall within 3σ of μ.

23

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…

24

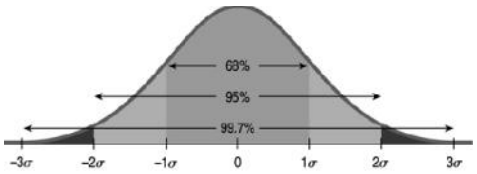
The 68-95-99.7 Rule (cont.)

- It turns out that in a Normal model:
  - about 68% of the values fall within one standard deviation of the mean;
  - about 95% of the values fall within two standard deviations of the mean; and,
  - about 99.7% (almost all!) of the values fall within three standard deviations of the mean.

25

The 68-95-99.7 Rule (cont.)

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



26

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.

27

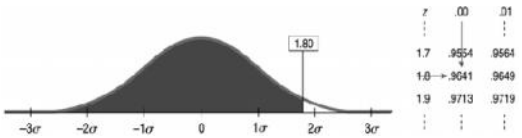
Finding Normal Percentiles by Hand

- 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.

28

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:



29

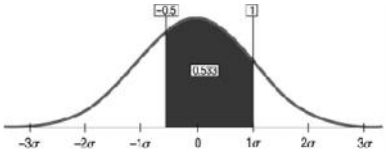
Finding Normal Percentiles Using Technology

- Many calculators and statistics programs have the ability to find normal percentiles for us.
- The *ActivStats* Multimedia Assistant offers two methods for finding normal percentiles:
  - The "Normal Model Tool" makes it easy to see how areas under parts of the Normal model correspond to particular cut points.
  - There is also a Normal table in which the picture of the normal model is interactive.

30

Finding Normal Percentiles Using Technology (cont.)

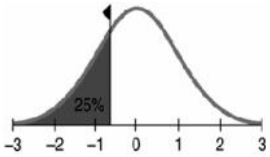
The following was produced with the "Normal Model Tool" in *ActivStats*:



31

From Percentiles to Scores: *z* in Reverse

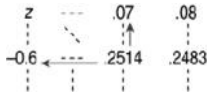
- Sometimes we start with areas and need to find the corresponding *z*-score or even the original data value.
- Example: What *z*-score represents the first quartile in a Normal model?



32

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.

33

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.

34

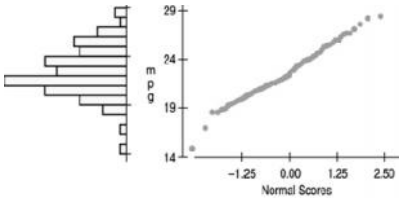
Are You Normal? How Can You Tell?

- 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.

35

Are You Normal? How Can You Tell?

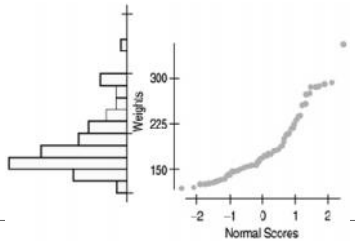
- **Nearly Normal** data have a histogram and a Normal probability plot that look somewhat like this example:



36

Are You Normal? How Can You Tell?

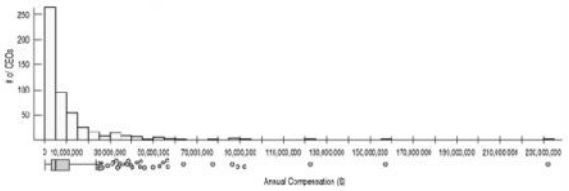
- A skewed distribution might have a histogram and Normal probability plot like this:



37

What Can Go Wrong?

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



38

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.

39

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.

40

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.

41

What have we learned? (cont.)

- We've learned that the 68-95-99.7 Rule can be a useful rule of thumb for understanding distributions:
  - For data that are unimodal and symmetric, about 68% fall within 1 SD of the mean, 95% fall within 2 SDs of the mean, and 99.7% fall within 3 SDs of the mean.

42

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?).

43

Percentiles

- Another measure of relative standing is a **percentile rank**.
- ***p*th percentile**: Value with ***p*** % of observations below it.
  - median = 50th percentile {*mean=50th %ile if symmetric*}
  - Q1 = 25th percentile
  - Q3 = 75th percentile

Jenny got an 86.
22 of the 25 scores are ≤ 86.
Jenny is in the 22/25 = 88th %ile.

6 | 7
7 | 2334
7 | 5777899
8 | 00123334
8 | 569
9 | 03

Percentiles

- Another measure of relative standing is a ***percentile rank***.
- ***p*th percentile**: Value with ***p*** % of observations below it.