Chapter 2: Modeling Distributions of Data

The Practice of Statistics, 4th edition - For AP*

STARNES, YATES, MOORE

Section 2.1
Describing Location in a Distribution

Learning Objectives

After this section, you should be able to…

- ✓ MEASURE position using percentiles
- ✓ INTERPRET cumulative relative frequency graphs
- ✓ MEASURE position using *z*-scores
- ✓ TRANSFORM data
- ✓ DEFINE and DESCRIBE density curves

Measuring Position: Percentiles

- One way to describe the location of a value in a distribution is to tell what percent of observations are less than it.

Definition:

The ***p*th percentile** of a distribution is the value with *p* percent of the observations less than it.

Example, p. 85

Jenny earned a score of 86 on her test. How did she perform relative to the rest of the class?

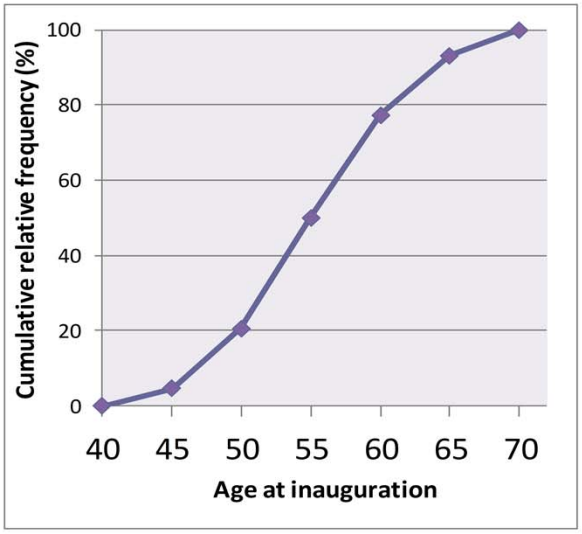
6	7
7	2334
7	5777899
8	00123334
8	569
9	03

Her score was greater than 21 of the 25 observations. Since 21 of the 25, or 84%, of the scores are below hers, Jenny is at the 84th percentile in the class's test score distribution.

Cumulative Relative Frequency Graphs

A **cumulative relative frequency** graph (or **ogive**) displays the cumulative relative frequency of each class of a frequency distribution.

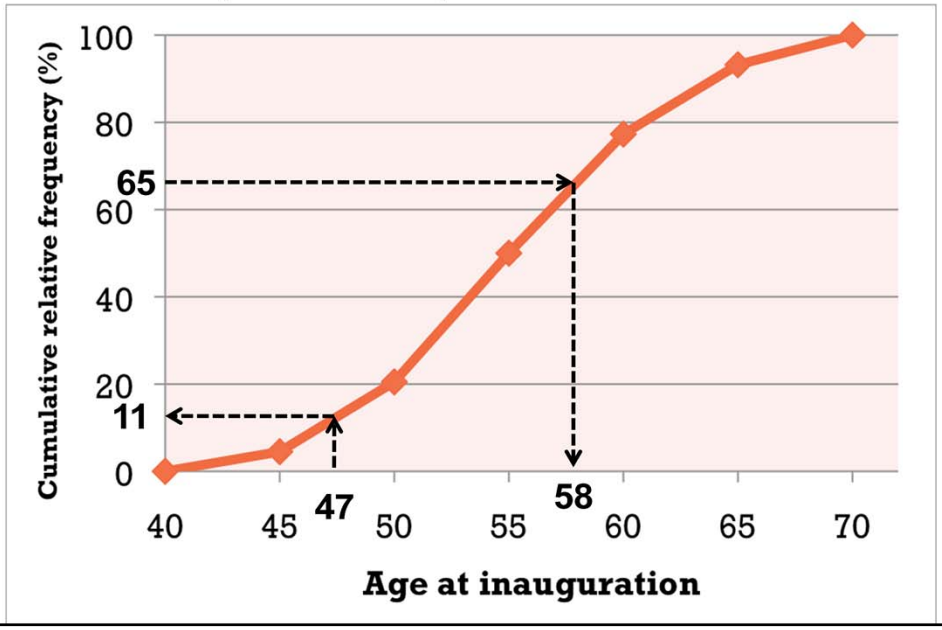
Age of First 44 Presidents When They Were Inaugurated				
Age	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
40-44	2	2/44 = 4.5%	2	2/44 = **4.5%**
45-49	7	7/44 = 15.9%	9	9/44 = **20.5%**
50-54	13	13/44 = 29.5%	22	22/44 = **50.0%**
55-59	12	12/44 = 34%	34	34/44 = **77.3%**
60-64	7	7/44 = 15.9%	41	41/44 = **93.2%**
65-69	3	3/44 = 6.8%	44	44/44 = **100%**



- **Interpreting Cumulative Relative Frequency Graphs**

Use the graph from page 88 to answer the following questions.

- Was Barack Obama, who was inaugurated at age 47, unusually young?
- Estimate and interpret the 65th percentile of the distribution



Measuring Position: *z*-Scores

- A *z*-score tells us how many standard deviations from the mean an observation falls, and in what direction.

Definition:

If *x* is an observation from a distribution that has known mean and standard deviation, the **standardized value** of *x* is:

$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$

A standardized value is often called a **z-score.**

Jenny earned a score of 86 on her test. The class mean is 80 and the standard deviation is 6.07. What is her standardized score?

$$z = \frac{x - \text{mean}}{\text{standard deviation}} = \frac{86 - 80}{6.07} = 0.99$$

Using *z*-scores for Comparison

We can use *z*-scores to compare the position of individuals in different distributions.

Example, p. 91

Jenny earned a score of 86 on her statistics test. The class mean was 80 and the standard deviation was 6.07. She earned a score of 82 on her chemistry test. The chemistry scores had a fairly symmetric distribution with a mean 76 and standard deviation of 4. On which test did Jenny perform better relative to the rest of her class?

$$z_{stats} = \frac{86-80}{6.07}$$

$$z_{stats} = 0.99$$

$$z_{chem} = \frac{82-76}{4}$$

$$z_{chem} = 1.5$$

Transforming Data

Transforming converts the original observations from the original units of measurements to another scale. Transformations can affect the shape, center, and spread of a distribution.

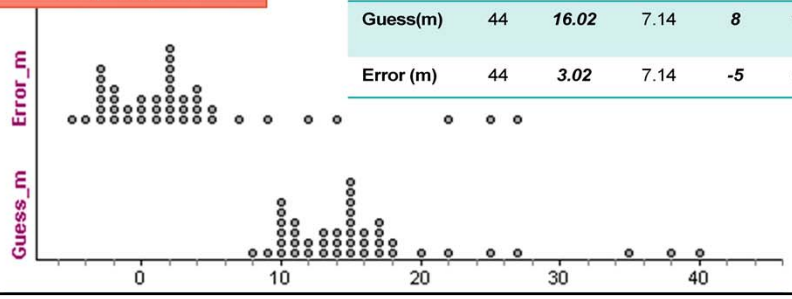
Effect of Adding (or Subracting) a Constant

Adding the same number *a* (either positive, zero, or negative) to each observation:

- adds *a* to measures of center and location (mean, median, quartiles, percentiles), but
- Does not change the shape of the distribution or measures of spread (range, *IQR*, standard deviation).

Example, p. 93

	n	Mean	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Guess(m)	44	***16.02***	7.14	***8***	***11***	***15***	***17***	***40***	6	32
Error (m)	44	***3.02***	7.14	***-5***	***-2***	***2***	***4***	***27***	6	32



Transforming Data

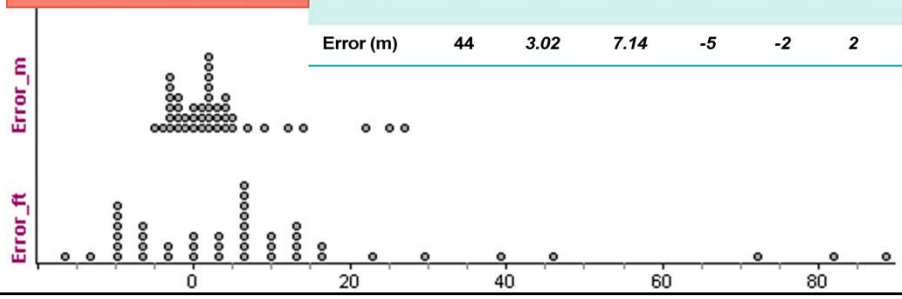
Effect of Multiplying (or Dividing) by a Constant

Multiplying (or dividing) each observation by the same number *b* (positive, negative, or zero):

- multiplies (divides) measures of center and location by *b*
- multiplies (divides) measures of spread by |*b*|, but
- does not change the shape of the distribution

Example, p. 95

	n	Mean	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Error(ft)	44	9.91	23.43	-16.4	-6.56	6.56	13.12	88.56	19.68	104.96
Error (m)	44	3.02	7.14	-5	-2	2	4	27	6	32



Density Curves

- In Chapter 1, we developed a kit of graphical and numerical tools for describing distributions. Now, we'll add one more step to the strategy.

Exploring Quantitative Data

1. Always plot your data: make a graph.
2. Look for the overall pattern (shape, center, and spread) and for striking departures such as outliers.
3. Calculate a numerical summary to briefly describe center and spread.
4. Sometimes the overall pattern of a large number of observations is so regular that we can describe it by a smooth curve.

Density Curve

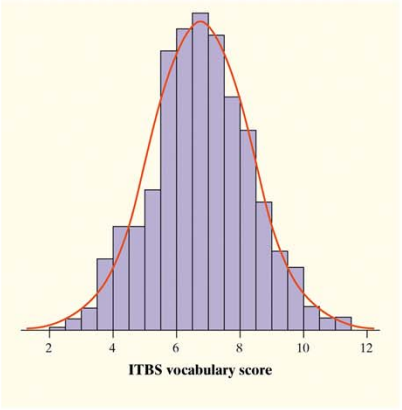
Definition:

A **density curve** is a curve that

- is always on or above the horizontal axis, and
- has area exactly 1 underneath it.

A density curve describes the overall pattern of a distribution. The area under the curve and above any interval of values on the horizontal axis is the proportion of all observations that fall in that interval.

The overall pattern of this histogram of the scores of all 947 seventh-grade students in Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills (ITBS) can be described by a smooth curve drawn through the tops of the bars.



Describing Density Curves

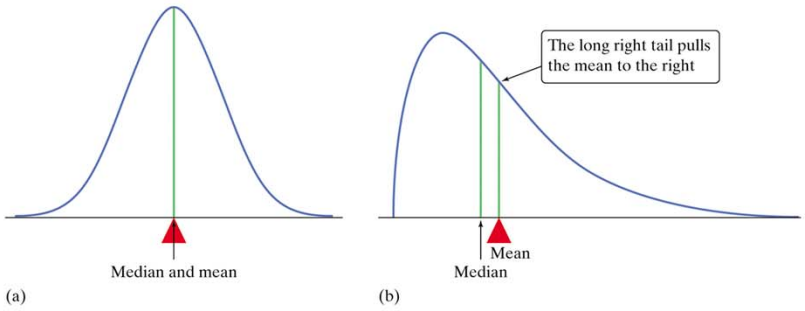
- Our measures of center and spread apply to density curves as well as to actual sets of observations.

Distinguishing the Median and Mean of a Density Curve

The **median** of a density curve is the equal-areas point, the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if made of solid material.

The median and the mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.



Section 2.2
Normal Distributions

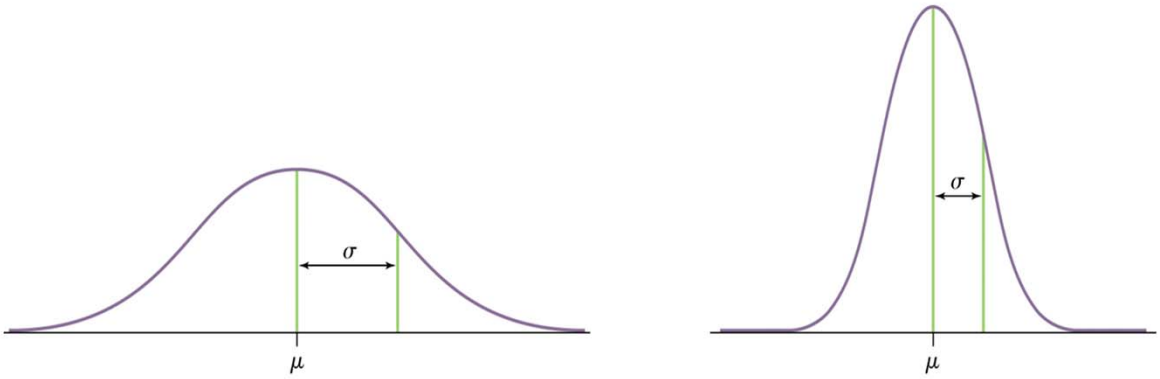
Learning Objectives

After this section, you should be able to…

- ✓ DESCRIBE and APPLY the 68-95-99.7 Rule
- ✓ DESCRIBE the standard Normal Distribution
- ✓ PERFORM Normal distribution calculations
- ✓ ASSESS Normality

Normal Distributions

- One particularly important class of density curves are the **Normal curves**, which describe **Normal distributions**.
  - All Normal curves are symmetric, single-peaked, and bell-shaped
  - A Specific Normal curve is described by giving its mean μ and standard deviation σ.



Two Normal curves, showing the mean μ and standard deviation σ.

Normal Distributions

Definition:

A **Normal distribution** is described by a Normal density curve. Any particular Normal distribution is completely specified by two numbers: its mean μ and standard deviation σ.

- The mean of a Normal distribution is the center of the symmetric **Normal curve**.
- The standard deviation is the distance from the center to the change-of-curvature points on either side.
- We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu,\sigma)$.

Normal distributions are good descriptions for some distributions of *real data.*

Normal distributions are good approximations of the results of many kinds of *chance outcomes*.

Many *statistical inference* procedures are based on Normal distributions.

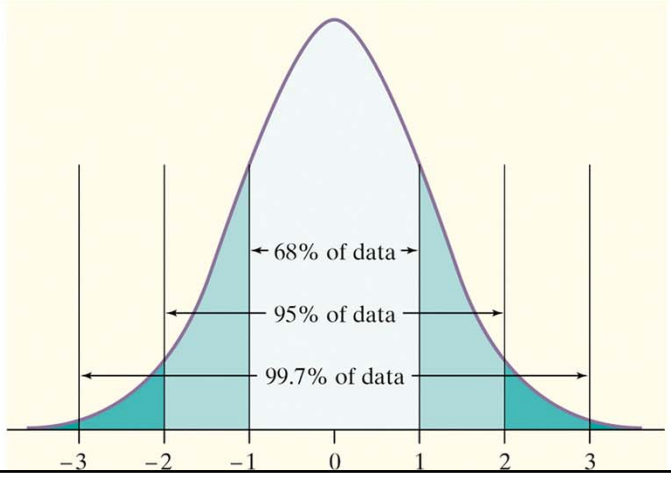
The 68-95-99.7 Rule

Although there are many Normal curves, they all have properties in common.

Definition: **The 68-95-99.7 Rule ("The Empirical Rule")**

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

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



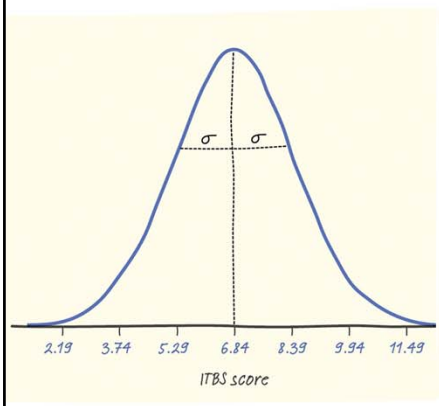
Example, p. 113

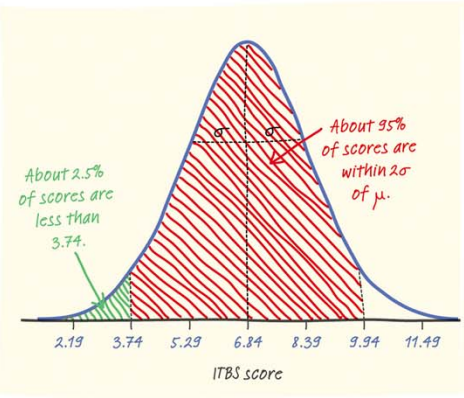
The distribution of Iowa Test of Basic Skills (ITBS) vocabulary scores for 7th grade students in Gary, Indiana, is close to Normal. Suppose the distribution is $N(6.84, 1.55)$.

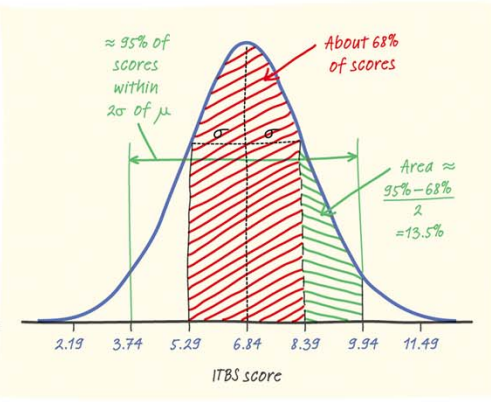
a) Sketch the Normal density curve for this distribution.

b) What percent of ITBS vocabulary scores are less than 3.74?

c) What percent of the scores are between 5.29 and 9.94?







The Standard Normal Distribution

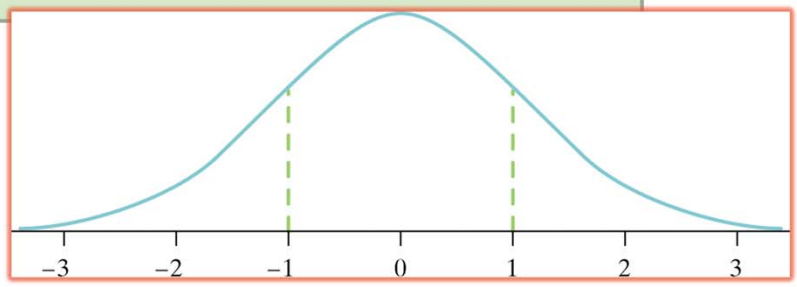
- All Normal distributions are the same if we measure in units of size σ from the mean μ as center.

Definition:

The **standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1.
If a variable x has any Normal distribution $N(\mu,\sigma)$ with mean μ and standard deviation σ, then the standardized variable

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

has the standard Normal distribution, $N(0,1)$.



The Standard Normal Table

Because all Normal distributions are the same when we standardize, we can find areas under any Normal curve from a single table.

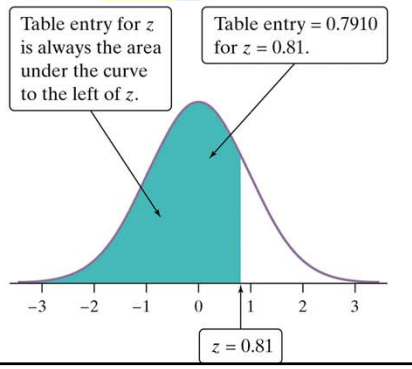
Definition: The Standard Normal Table

Table A is a table of areas under the standard Normal curve. The table entry for each value *z* is the area under the curve to the left of *z*.

Suppose we want to find the proportion of observations from the standard Normal distribution that are less than 0.81.
We can use Table A:

Z	.00	.01	.02
0.7	.7580	.7611	.7642
0.8	.7881	.7910	.7939
0.9	.8159	.8186	.8212

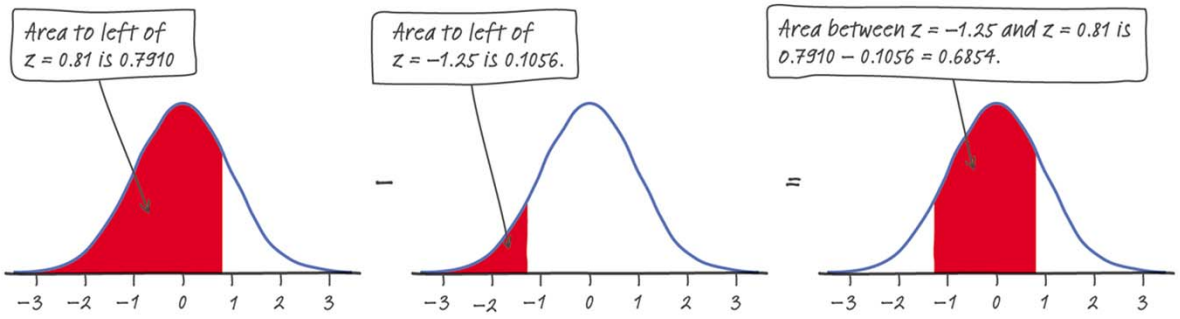
P(*z* < 0.81) = .7910



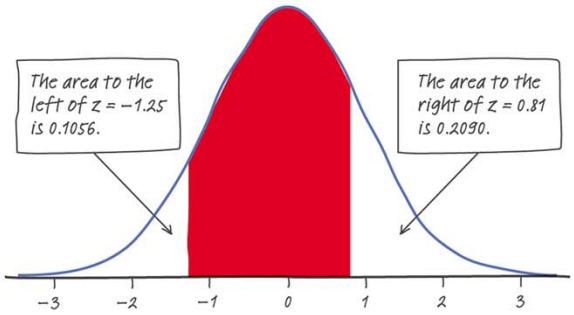
Example, p. 117

Finding Areas Under the Standard Normal Curve

Find the proportion of observations from the standard Normal distribution that are between -1.25 and 0.81.



Can you find the same proportion using a different approach?



$1 - (0.1056+0.2090) = 1 - 0.3146$

$= 0.6854$

Normal Distribution Calculations

How to Solve Problems Involving Normal Distributions

STEP 4

State: Express the problem in terms of the observed variable *x*.

Plan: Draw a picture of the distribution and shade the area of interest under the curve.

Do: Perform calculations.

- ***Standardize*** *x* to restate the problem in terms of a standard Normal variable *z*.
- ***Use Table A*** and the fact that the total area under the curve is 1 to find the required area under the standard Normal curve.

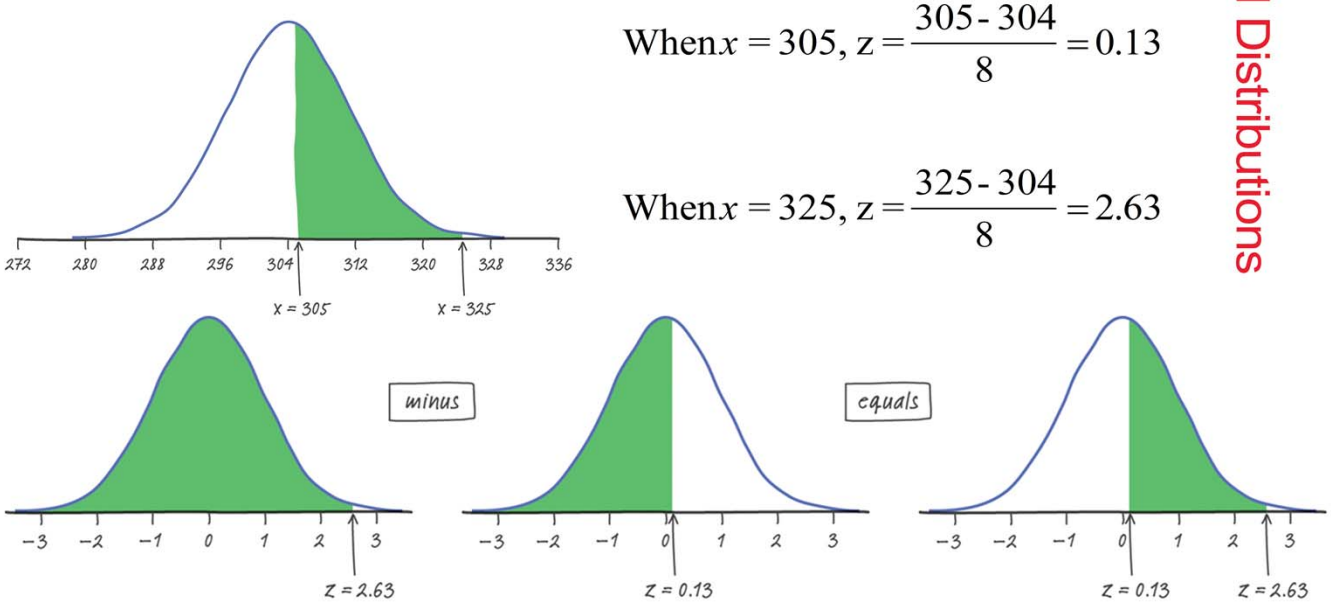
Conclude: Write your conclusion in the context of the problem.

Normal Distribution Calculations

When Tiger Woods hits his driver, the distance the ball travels can be described by *N*(304, 8). What percent of Tiger's drives travel between 305 and 325 yards?

$$\text{When } x = 305,\ z = \frac{305 - 304}{8} = 0.13$$

$$\text{When } x = 325,\ z = \frac{325 - 304}{8} = 2.63$$



Using Table A, we can find the area to the left of z=2.63 and the area to the left of z=0.13. **0.9957 – 0.5517 = 0.4440**. About **44%** of Tiger's drives travel between 305 and 325 yards.

Assessing Normality

The Normal distributions provide good models for some distributions of real data. Many statistical inference procedures are based on the assumption that the population is approximately Normally distributed. Consequently, we need a strategy for assessing Normality.

✓***Plot the data.***

- Make a dotplot, stemplot, or histogram and see if the graph is approximately symmetric and bell-shaped.

✓***Check whether the data follow the 68-95-99.7 rule.***

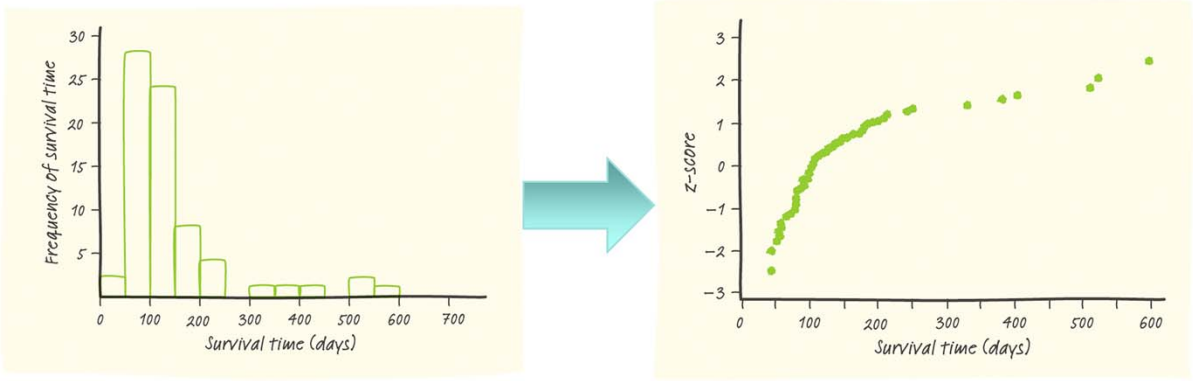
- Count how many observations fall within one, two, and three standard deviations of the mean and check to see if these percents are close to the 68%, 95%, and 99.7% targets for a Normal distribution.

Normal Probability Plots – Nice to Know but it will not be on AP Exam or any Test

- Most software packages can construct Normal probability plots. These plots are constructed by plotting each observation in a data set against its corresponding percentile's z-score.

Interpreting Normal Probability Plots

If the points on a **Normal probability plot** lie close to a straight line, the plot indicates that the data are Normal. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.



Summary

Section 2.1
Describing Location in a Distribution

In this section, we learned that…

- ✓ There are two ways of describing an individual's location within a distribution – the **percentile** and ***z*-score**.
- ✓ A **cumulative relative frequency graph** allows us to examine location within a distribution.
- ✓ It is common to **transform data**, especially when changing units of measurement. Transforming data can affect the shape, center, and spread of a distribution.
- ✓ We can sometimes describe the overall pattern of a distribution by a **density curve** (an idealized description of a distribution that smooths out the irregularities in the actual data).

Summary

Section 2.2
Normal Distributions

In this section, we learned that…

- ✓ The **Normal Distributions** are described by a special family of bell-shaped, symmetric density curves called **Normal curves**. The mean μ and standard deviation σ completely specify a Normal distribution $N(\mu,\sigma)$. The mean is the center of the curve, and σ is the distance from μ to the change-of-curvature points on either side.
- ✓ All Normal distributions obey the **68-95-99.7 Rule**, which describes what percent of observations lie within one, two, and three standard deviations of the mean.

Summary

Section 2.2
Normal Distributions

In this section, we learned that…

- ✓ All Normal distributions are the same when measurements are standardized. The **standard Normal distribution** has mean μ=0 and standard deviation σ=1.
- ✓ **Table A** gives percentiles for the standard Normal curve. By standardizing, we can use Table A to determine the percentile for a given *z*-score or the *z*-score corresponding to a given percentile in any Normal distribution.
- ✓ To assess Normality for a given set of data, we first observe its shape. We then check how well the data fits the **68-95-99.7 rule.** We can also construct and interpret a **Normal probability plot**.