Describing Location in a Distribution

Often, we are interested in describing where one observation falls in a distribution in relation to the other observations.

An individual’s **percentile** is the percent of values in a distribution that are less than the individual’s data value.

- Percentiles are specific *locations* – not regions – of a distribution, so an observation isn’t “in” the 80\(^{th}\) percentile. It’s “at” the 80\(^{th}\) percentile.
- Make sure you are referring to the observation (a number), not the individual. A person’s *height* is at the 30\(^{th}\) percentile, or the *volume* of soda in a bottle is at the 90\(^{th}\) percentile, not the person or the soda. Those aren’t numbers.

**Example:** The stemplot below shows the number of wins for each of the 30 Major League Baseball teams in 2018.

Find the percentiles of the number of wins for the following teams:

- **a)** The Los Angeles Dodgers, who won 92 games.
  
  The number of wins for the Dodgers was at the 80\(^{th}\) percentile in the distribution of win totals for MLB teams because \(24/30 = 80\%\) of teams won fewer games than they did in 2018.

- **b)** The Cincinnati Reds, who won 67 games.
  
  The number of wins for the Reds was at the 20\(^{th}\) percentile because \(6/20 = 20\%\) of teams won fewer games than they did in 2018. (Don’t count either of the 67s).

- **c)** The Boston Red Sox, who won 108 games.
  
  The number of wins for the Red Sox was at the 96.7\(^{th}\) percentile because \(29/30 \approx 96.7\%\) of teams won fewer games than they did in 2018.

- **d)** The number of wins for the Seattle Mariners is at the 60\(^{th}\) percentile of the distribution. How many games did they win?
  
  \[(0.6)(30) = 18\] teams won fewer games than the Mariners, so their win total is the 19\(^{th}\) one on the stemplot. They won 89 games in 2018.

**Note:** There is some disagreement over the exact definition of percentile. Some people define the \(p\)th percentile as the value *greater than or equal to* \(p\) percent of the observations. Any problems involving percentiles on the AP test will accept either definition.

**Cumulative Relative Frequency Graphs (Ogives)**

1. Start with a frequency table for a quantitative variable (similar to what you would use when making a histogram). Add columns for *relative frequency*, *cumulative frequency*, and *cumulative relative frequency*.
2. In the *relative frequency* column, divide each frequency by the total number of observations.
3. In the *cumulative frequency* column, add the counts in the frequency column for the current class and all classes with smaller values of the variable.
4. In the *cumulative relative frequency* column, divide the cumulative frequencies by the total number of observations.
5. Plot a point corresponding to the cumulative relative frequency in each class at the smallest value of the next class. Start with a point at a height of 0\% at the smallest value of the first class. The last point should have a height of 100\%. Connect consecutive points with line segments.
Example: The table shows the distribution of median household incomes for the 50 states and the District of Columbia. Complete the table and draw a cumulative relative frequency graph.

<table>
<thead>
<tr>
<th>Median Income ($1000s)</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Frequency</th>
<th>Cumulative Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>35 to &lt;40</td>
<td>1</td>
<td>1/51 = 0.0196</td>
<td>1</td>
<td>1/51 = 0.0196</td>
</tr>
<tr>
<td>40 to &lt;45</td>
<td>10</td>
<td>10/51 = 0.1961</td>
<td>11</td>
<td>11/51 = 0.2157</td>
</tr>
<tr>
<td>45 to &lt;50</td>
<td>14</td>
<td>14/51 = 0.2745</td>
<td>25</td>
<td>25/51 = 0.4902</td>
</tr>
<tr>
<td>50 to &lt;55</td>
<td>12</td>
<td>12/51 = 0.2353</td>
<td>37</td>
<td>37/51 = 0.7255</td>
</tr>
<tr>
<td>55 to &lt;60</td>
<td>5</td>
<td>5/51 = 0.0980</td>
<td>42</td>
<td>42/51 = 0.8235</td>
</tr>
<tr>
<td>60 to &lt;65</td>
<td>6</td>
<td>6/51 = 0.1176</td>
<td>48</td>
<td>48/51 = 0.9412</td>
</tr>
<tr>
<td>65 to &lt;70</td>
<td>3</td>
<td>3/51 = 0.0588</td>
<td>51</td>
<td>51/51 = 1</td>
</tr>
</tbody>
</table>

a) What does the point at (50,49) mean?
The point at (50,49) means that about 49% of states had median household incomes below $50,000.

b) At what percentile is California’s median household income of $57,445?
California’s median household income is at about the 78\textsuperscript{th} percentile. (About 78% of states had lower median household incomes than California.)

c) Estimate and interpret the first quartile (25\textsuperscript{th} percentile) of this distribution.
The first quartile of the distribution of median household incomes is about $46,000, which means that about 25% of states have median household incomes below $46,000.

d) What does the slope of the graph tell you?
The graph is steepest between 40 and 55, meaning that the cumulative relative frequency is increasing quickly. This means that there are lots of states with median household incomes between $40,000 and $55,000. The graph is less steep between 35 and 40 and above 55, meaning that there are fewer states with median incomes between $35,000 and $40,000 and above $55,000. There are lots of data points in intervals with steep slopes and very few data points in intervals with low slopes.
z-Scores

The **standardized score** or **z-score** of an observation tells us how many observations above or below the mean the value lies (above the mean if the z-score is positive, below the mean if the z-score is negative.)

\[
z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}
\]

**Example:** In 2018, the mean number of wins of the 30 Major League Baseball teams was 80.6 with a standard deviation of 14.5 wins. Find and interpret the z-scores for the following teams.

a) The Washington Nationals, who had 82 wins.

\[
z = \frac{82 - 80.6}{14.5} = 0.097
\]

The win total for the Nationals was 0.097 standard deviations above the mean for MLB teams in 2018.

b) The Baltimore Orioles, who had a dismal 47 wins.

\[
z = \frac{47 - 80.6}{14.5} = -2.32
\]

The win total for the Orioles was 2.32 standard deviations below the mean for MLB teams in 2018.

We often standardize values (use z-scores) to compare individuals in different distributions. In order for the comparison to be valid, it’s important that the distributions be roughly the same shape.

**Example:** The single-season home run record for Major League Baseball has been set just three times since Babe Ruth hit 60 home runs in 1927. Roger Maris hit 61 in 1961, Mark McGwire hit 70 in 1998, and Barry Bonds hit 73 in 2001. Baseball historians suggest that hitting a home run has been easier in some eras that others due to quality of hitters, quality of pitchers, hardness of the baseball, dimensions of the ballparks, and performance-enhancing drugs. To make a fair comparison, we should see how these performances rate relative to those of other hitters during the same year. Based on the information below, which player had the most outstanding performance relative to his peers?

Relative to his peers, Babe Ruth unquestionably had the best performance. In his record-breaking year, his HR total was 5.14 standard deviations above the mean for the year. This is higher than Maris, McGwire, and Bonds, whose HR totals were 3.15, 3.88, and 3.91 standard deviations above the means for their record-breaking years, respectively.

**Example:** Brent is a member of the school basketball team. He is 74 inches tall. The mean height of the players on the team is 76 inches. Brent’s height translates to a z-score of -0.85 in the team’s distribution of height. What is the standard deviation of heights for the members of the basketball team?

\[
-0.85 = \frac{74 - 76}{\sigma} \quad \Rightarrow \quad -0.85 = \frac{-2}{\sigma} \quad \Rightarrow \quad -0.85 \sigma = -2 \quad \Rightarrow \quad \sigma = 2.35 \text{ inches}
\]
Transforming Data

Often, we want to transform our data in some way, for example, if we want to convert the data to different units of measure. This might involve multiplying or dividing by a constant and/or adding or subtracting a constant. Standardizing (finding the z-scores) an entire distribution is another example of transforming data.

Example: Here are a graph and summary statistics for a sample of 30 test scores. The maximum possible score on the test was 50 points.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Count</th>
<th>Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
<th>Range</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>30</td>
<td>35.80</td>
<td>8.17</td>
<td>12.00</td>
<td>32.00</td>
<td>37.00</td>
<td>41.25</td>
<td>48.00</td>
<td>36.00</td>
<td>9.25</td>
</tr>
</tbody>
</table>

a) Suppose the teacher was nice and added 5 points to each test score. What effect would this have on the shape of the distribution and measures of location, center, and variability?

Measures of variability (range, IQR, standard deviation): No change – the points aren’t any more or less spread out than they were before. They just shifted along a number line.

b) Suppose the teacher wants to convert the original test scores to percents by multiplying each score by 2. What effect would this have on the shape of the distribution and measures of location, center, and variability?

Measures of variability (range, IQR, standard deviation): All multiplied by 2 – the points are twice as far apart as they were before.

Effect of Adding (or Subtracting) a Constant

Adding the same number \(a\) (positive, negative, or zero) to each observation

- adds \(a\) to the measures of center and location (mean, median, quartiles, percentiles).
- does not change the shape of the distribution or the measures of variability (range, IQR, standard deviation).

Effect of Multiplying (or Dividing) a Constant

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

- multiplies (divides) measures of center and location (mean, median, quartiles, percentiles) by \(b\).
- multiplies (divides) measures of variability (range, IQR, standard deviation) by \(|b|\).
- does not change the shape of the distribution.
Example: Knoebels Amusement Park in Elysburg, Pennsylvania has earned acclaim for being an affordable, family-friendly entertainment venue. Knoebels does not charge for general admission or parking, but it does charge customers for each ride they take. How much do the rides cost at Knoebels? The figure below shows a dotplot of the cost for each of the 22 rides in a recent year, along with summary statistics.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost ($)</td>
<td>22</td>
<td>1.705</td>
<td>0.447</td>
<td>1.25</td>
<td>1.5</td>
<td>1.5</td>
<td>1.75</td>
<td>2.75</td>
<td>3</td>
</tr>
</tbody>
</table>

a) Suppose you convert the cost of the rides from dollars to cents. Describe the shape, mean, and standard deviation of the distribution of ride cost in cents. Change: Multiply dollar amounts by 100.

Shape: Same. The distribution of ride costs in cents would be unimodal and skewed to the right. Lower costs are more common than higher costs. The ride costing 300¢ is a possible high outlier.

Mean = 100 (1.705) = 170.5¢  
St. Dev. = 100 (0.447) = 44.7¢

b) Knobels’ managers decide to increase the cost of each ride by 25 cents. How would the shape, center, and variability of this distribution compare with the original distribution of cost? Change: Add $0.25.

Shape: Same

Mean = $1.705 + $0.25 = $1.925  
St. Dev = $0.447  
(Adding affects the mean, but not the standard deviation.)

c) Suppose you convert the original costs to z-scores. What would be the shape, mean, and standard deviation of this distribution?

Subtracting affects the mean but not the standard deviation. Dividing affects both.

Mean = \( \frac{1.705 - 1.705}{0.447} = 0 \)  
St. Dev. = \( \frac{0.447}{0.447} = 1 \)

Distributions of z-scores always have mean 0 and standard deviation 1. This makes sense, because z-scores tell you the number of standard deviations above or below the mean for each observation in a distribution.

Example: In 2010, taxi cabs in New York City charged an initial fee of $2.50 plus $2 per mile, or in equation form, fare = 2.5 + 2(miles). At the end of a month, a businessman collects all his taxi cab receipts and calculates some numerical summaries. The mean fare he paid was $15.45, with a standard deviation of $10.20. What are the mean and standard deviation of the distances traveled on his cab rides, in miles?

Solve the equation for miles: miles = \( \frac{\text{fare} - 2.50}{2} \). (Subtract 2.50, then divide by 2).

Subtracting changes the mean, but not the standard deviation. Dividing changes both.

Mean = \( \frac{15.45 - 2.5}{2} = 6.475 \) miles  
St. Dev. = \( \frac{10.20}{2} = 5.10 \) miles
Density Curves and Normal Distributions

“All models are wrong, but some are useful” – George E. P. Box

Density Curves
Sometimes, the overall shape of a distribution of a large number of observations is so regular that we can describe it by a smooth curve.

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 gives the proportion of all observations that fall in that interval.

Density curves come in many shapes. Outliers are not described by a density curve. Density curves are approximations of real data. They aren’t perfect, but they simplify things a lot.

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, the point at which the curve would balance if made of solid material.

The median and 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.

Since density curves are idealized descriptions of data, we use the Greek letters $\mu$ (mu) and $\sigma$ (sigma) to represent the mean and standard deviation, respectively, of a density curve. These numbers are called parameters.

★ For most distributions, just knowing $\mu$ and $\sigma$ is not enough to tell you about the appearance of the distribution or to allow you to answer questions about percentiles or the area under the curve. You also need information about the shape of the distribution. However, there is a special type of curve called a Normal curve that shows up often in statistics. If you know you are dealing with an approximately Normal distribution, the only other information you need to know to answer all sorts of questions is the mean and standard deviation of the distribution.
Normal Curves

- **Shape:** All normal curves have the same overall shape: symmetric, unimodal, and bell-shaped.
- **Center:** The mean, $\mu$, of a Normal distribution is located at the center of the symmetric density curve and is the same as the median.
- **Variability:** The standard deviation, $\sigma$, of a Normal distribution is the distance from the center of the curve to the change-of-curvature points (inflection points) on either side. These are the places where the curve changes from getting steeper to flattening out. (If you have trouble seeing it, the point where this occurs is about 60% of the way up from the $x$-axis to the top of the curve. You can think of it as the point where the curve changes from making a “sad” face to making a “happy” face.)
- Any specific Normal curve is completely described by giving its mean $\mu$ and standard deviation $\sigma$. We abbreviate the Normal distribution with mean $\mu$ and standard deviation $\sigma$ as $N(\mu, \sigma)$.

Note: The equation of a Normal curve is $f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Any former calculus students are welcome to try to find areas under the curve by integrating this function rather than using a table or calculator. Have fun with that!

Why are Normal distributions important?

- They are good descriptions for some distributions of real data, like standardized test scores, repeated careful measurements of the same quantity, and characteristics of biological populations (lengths of crickets, corn yields, human heights, etc.)
- They are good approximations to the results of many kinds of chance outcomes, like the number of heads in many tosses of a fair coin.
- Many statistical inference procedures are based on Normal distributions.

★ Remember, not all data sets follow a Normal distribution! It is important to check to see whether the data are Normally distributed before continuing with many statistical procedures.

The 68-95-99.7 Rule (or the Empirical Rule)

In a Normal distribution:

- Approximately **68%** of the observations fall within **1 standard deviation** of the mean.
- Approximately **95%** of the observations fall within **2 standard deviations** of the mean.
- Approximately **99.7%** of the observations fall within **3 standard deviations** of the mean.

★ THIS RULE ONLY APPLIES TO NORMAL DISTRIBUTIONS!
How to draw a beautiful Normal curve:
1. Draw a number line with seven evenly spaced tick marks. Label the one in the middle with the mean value, $\mu$. The others represent the points 1, 2, and 3 standard deviations away from the mean.
2. Draw a dot above the mean value where you want the peak of the curve to be.
3. Above the marks one standard deviation from the mean, draw dots about 60% as high as the dot above the mean.
4. Above the marks two standard deviations from the mean, draw dots about 25% as high as the dots you just drew (or about 15% as high as the dot above the mean).
5. Above the marks three standard deviations from the mean, draw dots just barely above the axis.
6. Connect the seven dots with a smooth curve.

Example: The weights of yearling Angus steers are approximately Normally distributed with a mean of 1152 pounds and a standard deviation of 84 pounds.

- Sketch a Normal density curve for this distribution of steer weights. Label the points that are 1, 2, and 3 standard deviations from the mean.
- What percent of the steers weigh more than 1320 lbs.? Show your work.
- What percent of the steers weigh between 984 and 1236 lbs.? Show your work.

b) 1320 lbs. is 2 standard deviations above the mean. We know that about 95% of steer weights were between 984 and 1320 lbs. About half of the remaining 5%, or about 2.5% were above 1320 lbs.

c) About 68% of batting averages were between 1068 lbs. and 1236 lbs. Also, half of the difference between 95% and 68% were between 984 lbs. and 1068 lbs. Therefore, about 13.5% + 68% = 81.5% were between 984 lbs. and 1236 lbs.

Chebyshev’s Inequality
The Empirical Rule applies only to Normal distributions, but there is a more general rule that applies to any distribution. It is called Chebyshev’s inequality:

- In any distribution, the proportion of observations falling within $k$ standard deviations of the mean is at least $1 - \frac{1}{k^2}$.

Standard Normal distribution
The standard Normal distribution is the Normal distribution with mean 0 and standard deviation 1. In other words, it is a distribution of $z$-scores.

If a variable $x$ has any Normal distribution with $\mu = 0$ and $\sigma = 1$, then the standardized variable $z = \frac{x-\mu}{\sigma}$ has the standard Normal distribution. (In other words, if some poor sucker spent way too much time finding the areas under a standard Normal curve, then we can find the areas under any Normal curve using $z$-scores).
Standard Normal table
The standard Normal table shows $z$-scores in the margins (tenths along the left and hundredths along the top). The entries in the table are the areas under the standard Normal curve to the left of $z$.

For example, to find the area under the curve to the left of $z = 0.81$, find the row for 0.8 and the column for 0.01 and find the corresponding table entry, which is 0.7910.

<table>
<thead>
<tr>
<th>$z$</th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>.7580</td>
<td>.7611</td>
<td>.7642</td>
</tr>
<tr>
<td>0.8</td>
<td>.7881</td>
<td>.7910</td>
<td>.7939</td>
</tr>
<tr>
<td>0.9</td>
<td>.8159</td>
<td>.8186</td>
<td>.8212</td>
</tr>
</tbody>
</table>

Normal curve calculations on TI-83/TI-84 calculators:
- If you know values from the distribution or $z$-scores and want to find the area under the curve:
  - Press 2ND VARS to get to the DISTR menu.
  - Choose the command normalcdf(lower bound, upper bound, $[\mu, \sigma]$)
    - Entering $\mu$ and $\sigma$ is optional. If you are using $z$-scores, either don’t enter them or leave them as $\mu = 0$, $\sigma = 1$.
    - When finding areas in the tails of the curve, use very large positive and negative numbers, like $\pm 100000$ or $\pm 1 \times 10^9$ for $\pm \infty$ as your lower and upper bounds.
- If you know the area under the curve and want a $z$-score or a value from the distribution:
  - Press 2ND VARS to get to the DISTR menu.
  - Choose the command invNorm(area to the left of the boundary, $[\mu, \sigma]$).
    - Entering $\mu$ and $\sigma$ is optional. If you don’t enter them or leave them as $\mu = 0$, $\sigma = 1$, the command will give you the $z$-score corresponding to the boundary value. If you enter a specific $\mu$ and $\sigma$, the command will give you the actual boundary value from the distribution.

**Examples:** Find the proportion of observations from a standard Normal distribution that fall in each of the following regions. For each example, sketch a standard Normal curve and shade the area representing the region.

a) $z < -1.23$

Area = 0.1093

$z = -1.23$

$0.1093 = \text{normalcdf (low = -1E99, high = -1.23)}$

b) $z > 2.47$

Area = 0.0068

$z = 2.47$

$1 - 0.9932 = 0.0068 = \text{normalcdf (low = 2.47, high = 1E99)}$

c) $-0.32 < z < 1.58$

Area = 0.5684

$z = -0.32$

$z = 1.58$

$0.9429 - 0.3745 = 0.5684 = \text{normalcdf (low = -0.32, high = 1.58)}$

Find the value of $z$ from the standard Normal distribution that satisfies each of the following conditions. In each case, sketch a standard Normal curve with your value of $z$ marked on the axis.

a) The 30th percentile

$z = -0.5244$

$\text{invNorm (area to left = 0.3)}$

b) 5% of all values are greater than $z$

$z = 1.645$

$\text{invNorm (area to left = 0.95)}$ (Enter area to the left of $z$)
Normal Curve Problems

Draw a Normal curve. Label the $x$-axis. Label $\mu$ and $\sigma$ (with values if known, symbols if not).

<table>
<thead>
<tr>
<th>If question asks about percent, proportion, probability, or percentile:</th>
<th>If question asks about a boundary value, mean, or standard deviation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Answer is the area of a region under the curve.</td>
<td>• Use the invNorm command to find a $z$-score corresponding to a known area, then use the $z$-score formula to solve for the unknown value.</td>
</tr>
<tr>
<td>1. Calculate the $z$-score(s) of the boundary value(s) using the $z$-score formula: $z = \frac{\text{value} - \mu}{\sigma}$. (Although the calculator can find the final answer without using $z$-scores, always show this calculation anyway! It doesn’t take that long and the AP rubrics have been unpredictable lately, so I am requiring it.)</td>
<td>1. Draw the boundary on your normal curve with the correct area of interest shaded. Label the area. Label the boundary value with a variable (if unknown) or its value (if known).</td>
</tr>
<tr>
<td>2. Draw the boundary/boundaries on your normal curve the correct number of standard deviations above or below the mean based on the $z$-score(s). Label the boundary value(s). Shade the area of interest under the curve.</td>
<td>2. Use invNorm command to find the $z$-score corresponding to the boundary value for the known area. $\text{invNorm(} \text{area to left of boundary, } [\mu = 0, \sigma = 1] \text{)}$.</td>
</tr>
<tr>
<td>3. Use the normalcdf command to find the area of the shaded region under the curve. $\text{normalcdf(} \text{lowerbound, upperbound, } [\mu, \sigma] \text{)}$</td>
<td>3. Plug the $z$-score and known values into the $z$-score formula to solve for the unknown value: $z = \frac{\text{value} - \mu}{\sigma}$.</td>
</tr>
</tbody>
</table>

If entering $z$-scores for the boundary values, leave $\mu = 0$ and $\sigma = 1$ or don’t enter a mean and standard deviation.

If entering the actual boundary values instead of their $z$-scores, enter the mean and standard deviation for $\mu$ and $\sigma$ instead of $\mu = 0$ and $\sigma = 1$. (The $z$-score calculations still need to be written down, even if you use this option!)

Note: When finding areas in the tails of the curve, use very large positive and negative numbers, like $\pm 100000$ or $\pm 1E99$ for $\pm \infty$. Note: If it is the boundary value that is unknown, the calculator can bypass the equation step. Enter the mean and standard deviation instead of entering $\mu = 0$ and $\sigma = 1$. The output will be the actual boundary value rather than the $z$-score of the boundary value. Last year, an AP rubric dropped scores from E to I if students didn’t use an equation, so you can only do this for a check – note that your answer may be slightly off due to rounding errors. If it is the mean or standard deviation that is unknown, there is no shortcut!
Example: In the 2008 Wimbledon tennis tournament, Rafael Nadal averaged a serve speed of 115 mph on his first serves. Assume that the distribution of his first-serve speeds is approximately Normal with $\mu = 115$ mph and $\sigma = 6$ mph.

a) About what proportion of Nadal’s first serves would you expect to exceed 120 mph?

Let $x$ = the speed of Nadal’s first serve.

For what proportion of serves is $x > 120$?

$$z = \frac{120 - 115}{6} = 0.83$$

$normalcdf$ (lowerbound = 0.83, upperbound = 1E99)

$P(x > 120) = P(z > 0.83) = 0.2033$

Approximately 20.3% of Nadal’s first serves exceed 120 mph.

b) About what percent of Nadal’s first serves were between 100 and 110 mph?

For what proportion of serves is $100 < x < 110$?

$$z = \frac{100 - 115}{6} = -2.5 \\ z = \frac{110 - 115}{6} = -0.83$$

$normalcdf$ (lowerbound = -2.5, upperbound = -0.83)

$P(100 < x < 110) = P(-2.5 < z < -0.83) = 0.1971$

Approximately 19.7% of Nadal’s first serves were between 100 & 110 mph.

Example: According to the CDC, the heights of 3-year-old girls are approximately Normally distributed with a mean of 94.5 cm and a standard deviation of 4 cm.

a) What is the third quartile of the distribution?

What height is taller than 75% of 3-year-old girls?

$invNorm$ (area to left = 0.75). This gives the $z$-score for the third quartile ($z \approx 0.675$).

$$0.675 = \frac{x - 94.5}{4}$$

$$2.7 = x - 94.5$$

$$x = 97.2 \text{ cm}$$

The third quartile of heights for 3-year-old girls is about 97.2 cm. That is, about 75% of 3-year-old girls are shorter than 97.2 inches.

a) A particular 3-year-old girl is shorter than 95% of 3-year-old girls. What is her height?

What height is shorter than 95% of 3-year-old girls?

$invNorm$ (area to left = 0.05). This gives the $z$-score for this value ($z = -1.645$).

$$-1.645 = \frac{x - 94.5}{4}$$

$$-6.58 = x - 94.5$$

$$x = 87.92 \text{ cm}$$

A 3-year-old girl who is shorter than 95% of 3-year-old girls has a height of about 87.9 cm.
Example: Agricultural scientists are working on developing an improved variety of Roma tomatoes. Marketing research indicates that customers are likely to bypass Romas that weigh less than 70 grams. The current variety of Roma plants produces fruit that averages 74 grams, but 11% of the tomatoes are too small. It is reasonable to assume that a Normal model applies.

a) What is the standard deviation of the weights of Romas now being grown?

We need to find the \( z \)-score that corresponds to 70 grams. We do this by using the fact that the area to the left of 70 is 0.11. So, we use the command `invNorm (area = 0.11)` to get the \( z \)-score. This gives \( z \approx -1.2265 \).

\[
\begin{align*}
70 & - 74 \\
\sigma & = 3.3 \\
-1.2265 & = \frac{70 - 74}{\sigma} \\
-1.2265\sigma & = -4 \\
\sigma & \approx 3.3 \text{ grams}
\end{align*}
\]

The standard deviation of the tomato weights is about 3.3 grams.

b) Scientists hope to reduce the frequency of undersized tomatoes to no more than 4%. One way to accomplish this is to raise the average size of the fruit. If the standard deviation remains the same, what target mean should they have as a goal?

We need to find the \( z \)-score that corresponds to 70 grams again. This time, the area to the left of 70 is 0.04. So, we use the command `invNorm (area = 0.04)` to get the \( z \)-score. This gives \( z \approx -1.75 \).

\[
\begin{align*}
70 & - \mu \\
3.3 & = 70 - \mu \\
-5.775 & = 70 - \mu \\
\mu & \approx 75.8 \text{ grams}
\end{align*}
\]

The scientists should have a target mean weight of about 75.8 grams if they want only 4% of the tomatoes to be undersized.

Assessing Normality

Since many statistical procedures only work for distributions that are close to Normal, it’s important to have methods for checking whether a Normal distribution is a reasonable model for our observed data. One strategy is to plot the data, look at the shape of the distribution, and check to see whether it follows the 68-95-99.7 rule.

★ No distribution of real-world data will ever be exactly Normally distributed! The real world is messy. Make sure to say “approximately” Normal when describing distributions of real-world data.
Example: The measurements listed below describe the usable capacity (in cubic feet) of a sample of 36 side-by-side refrigerators. Is the distribution of refrigerator capacities approximately Normal?

<table>
<thead>
<tr>
<th>Capacity (cubic feet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.9</td>
</tr>
<tr>
<td>13.7</td>
</tr>
<tr>
<td>14.1</td>
</tr>
<tr>
<td>14.2</td>
</tr>
<tr>
<td>14.5</td>
</tr>
<tr>
<td>14.5</td>
</tr>
<tr>
<td>14.6</td>
</tr>
<tr>
<td>14.7</td>
</tr>
<tr>
<td>15.1</td>
</tr>
<tr>
<td>15.2</td>
</tr>
<tr>
<td>15.3</td>
</tr>
<tr>
<td>15.3</td>
</tr>
<tr>
<td>15.3</td>
</tr>
<tr>
<td>15.3</td>
</tr>
<tr>
<td>15.3</td>
</tr>
<tr>
<td>15.5</td>
</tr>
<tr>
<td>15.6</td>
</tr>
<tr>
<td>15.6</td>
</tr>
<tr>
<td>15.6</td>
</tr>
<tr>
<td>15.8</td>
</tr>
<tr>
<td>16.0</td>
</tr>
<tr>
<td>16.0</td>
</tr>
<tr>
<td>16.2</td>
</tr>
<tr>
<td>16.2</td>
</tr>
<tr>
<td>16.3</td>
</tr>
<tr>
<td>16.4</td>
</tr>
<tr>
<td>16.5</td>
</tr>
<tr>
<td>16.6</td>
</tr>
<tr>
<td>16.6</td>
</tr>
<tr>
<td>16.6</td>
</tr>
<tr>
<td>16.8</td>
</tr>
<tr>
<td>17.0</td>
</tr>
<tr>
<td>17.0</td>
</tr>
<tr>
<td>17.2</td>
</tr>
<tr>
<td>17.4</td>
</tr>
<tr>
<td>17.4</td>
</tr>
<tr>
<td>17.9</td>
</tr>
<tr>
<td>18.4</td>
</tr>
</tbody>
</table>

a) Display the data graphically to make sure that the distribution is unimodal, approximately symmetric, and bell-shaped.

![Histogram of Refrigerator Capacity](image)

The distribution of refrigerator capacity is unimodal and reasonably symmetric. It does appear fairly bell-shaped.

b) Calculate the mean and standard deviation of the data. Determine what percent of the observations fall within one, two, and three standard deviations of the mean. Compare your results to the 68-95-99.7 rule.

\[ \bar{x} = 15.825 \text{ ft}^3 \quad s_x = 1.217 \text{ ft}^3 \]

\[ \bar{x} \pm 1s_x = 15.825 \pm 1.217 = (14.608, 17.042) \]

24 of 36 refrigerators fall into this interval: 66.7%

\[ \bar{x} + 2s_x = 15.825 \pm 2(1.217) = (13.391, 18.259) \]

34 of 36 refrigerators fall into this interval: 94.4%

\[ \bar{x} + 3s_x = 15.825 \pm 3(1.217) = (12.174, 19.476) \]

36 of 36 refrigerators fall into this interval: 100%

These percentages are close to what we would expect based on the 68-95-99.7 rule. 66.7% of refrigerator capacities are within 1 standard deviation of the mean, 94.4% are within 2 standard deviations of the mean, and 100% are within 3 standard deviations of the mean. Combined with the histogram, this is good evidence that the distribution of refrigerator capacities is close to Normal.

Normal Probability Plots

A Normal probability plot is a scatterplot of the ordered pair (data value, expected z-score) for each of the individuals in a quantitative data set. That is, the x-coordinate of each point is the actual data value and the y-coordinate is the expected z-score corresponding to the percentile of that data value in a standard Normal distribution.

Normal probability plots are rarely made by hand, so you don’t need to worry about the details. Just know what you are looking for.
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 approximately Normal. A non-linear form (curved patterns, etc.) in a Normal probability plot indicates a non-Normal distribution.

When you look at a Normal probability plot, look for shapes that show clear departures from Normality. Don’t overreact to minor wiggles in the plot.

Example: Make a Normal probability plot for the refrigerator data in the previous example. Sketch your result. What do you observe?

The Normal probability plot is very linear. This is more good evidence that the distribution of refrigerator capacities is approximately Normal.

Example: Here are a histogram and Normal probability plot of NBA free-throw percentages. Based on the graphs, do free-throw percentages appear to be Normally-distributed?

Free-throw percentages do not appear to be Normally-distributed. The histogram shows that the distribution is skewed to the left, and the Normal probability plot has a curved pattern.