TOPIC 2.1

Frequency Distributions and Their Graphs

OBJECTIVES Build a frequency distribution with limits, midpoints, relative frequencies, and cumulative frequencies. Draw a frequency histogram, a frequency polygon, and an ogive.

GETTING STARTED Babe Ruth hit these home run totals for the Yankees from 1920 to 1934 (Source: Baseball-Reference):

54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22

a.) Put the data in order from least to greatest.

22, 25, 34, 35, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60

b.) In a typical season, about how many home runs did Ruth hit? Why is that hard to answer from the unsorted list?

About 45 to 50. The unsorted list hides the pattern; grouping the data makes it easy to see.

TIP
Let students guess first. The rest of the lesson organizes the data so the answer is obvious.

Frequency Distributions

DEFINITION **Frequency Distribution**

A **frequency distribution** is a table that shows **classes** (intervals) of data entries with a count of the entries in each class. The count is the **frequency**, f.

Each class has a **lower class limit** (least number that can belong to the class) and an **upper class limit** (greatest number).

The **class width** is the distance between the lower limits of consecutive classes.
The **range** is the maximum entry minus the minimum entry.

PROCEDURE Building a Frequency Distribution

1. Choose the number of classes, usually 5 to 20.
2. Class width = $\frac{\text{range}}{\text{number of classes}}$, rounded **up** to the next convenient number.
3. The minimum entry is the first lower limit. Add the class width to get each next lower limit.
4. Each upper limit is one less than the next lower limit.
5. Tally the entries and count the frequency of each class.

DEFINITION More Columns

Midpoint = $\frac{\text{lower limit+upper limit}}{2}$

Relative frequency = $\frac{f}{n}$, where n is the total number of entries.

Cumulative frequency of a class is the sum of the frequency of that class and all previous classes.

Example 1 Build a frequency distribution of Ruth's home runs with 5 classes.

a.) Find the range. 38

range = max − min

range = 60 − 22 = 38

b.) Find the class width. 8

width = $\frac{\text{range}}{\text{classes}} = \frac{38}{5} = 7.6 \rightarrow$ round up: 8

TIP
Round UP even when the decimal is small, or the last entry may not fit.

c.) Complete the table.

Class	Frequency, f	Midpoint	Relative frequency	Cumulative frequency
22–29	2	25.5	0.13	2
30–37	2	33.5	0.13	4
38–45	2	41.5	0.13	6
46–53	5	49.5	0.33	11
54–61	4	57.5	0.27	15

midpoint $= \frac{22+29}{2} = 25.5$ rel. freq. $= \frac{2}{15} \approx 0.13$

check: $2+2+2+5+4 = 15 = n$ last cumulative frequency $= n$

d.) In what percent of these seasons did Ruth hit 54 or more home runs? about 27%

$\frac{4}{15} \approx 0.27 = 27\%$

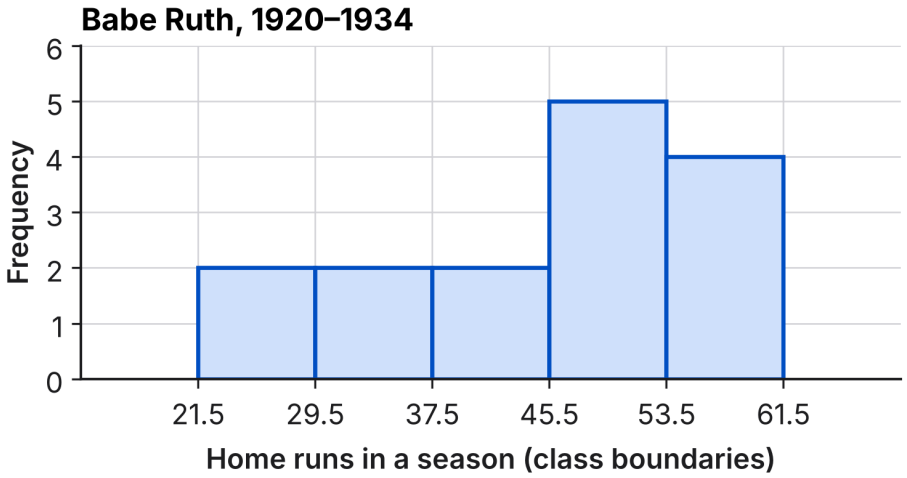
Graphs of Frequency Distributions

DEFINITION **Frequency Histogram**

A **frequency histogram** is a bar graph whose horizontal scale is quantitative. Its bars touch .

The bars run between **class boundaries**: the numbers halfway between the upper limit of one class and the lower limit of the next. For example, the boundary between 22–29 and 30–37 is 29.5 .

Example 2 Draw a frequency histogram of Ruth's home runs.

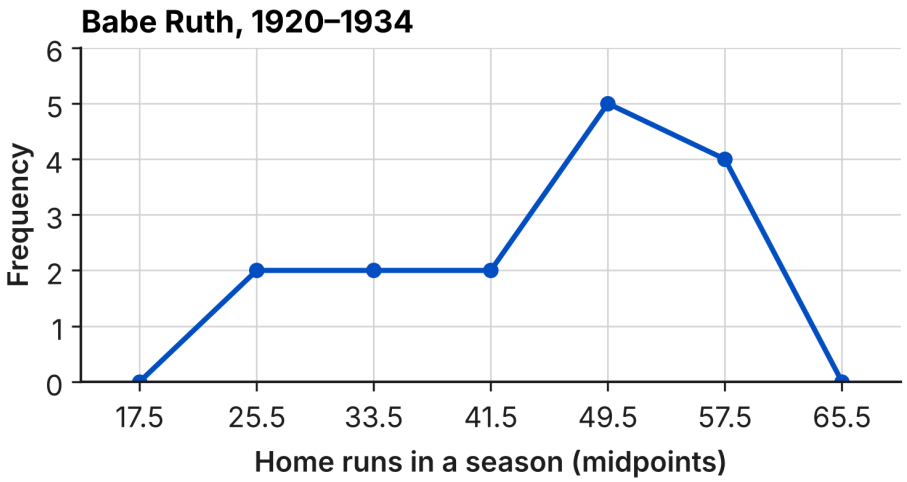


DEFINITION **Frequency Polygon and Relative Frequency Histogram**

A **frequency polygon** plots each class midpoint against its frequency and connects the points. It starts and ends on the horizontal axis, one class width beyond the first and last midpoints.

A **relative frequency histogram** has the same shape as the frequency histogram, but its vertical scale shows relative frequencies.

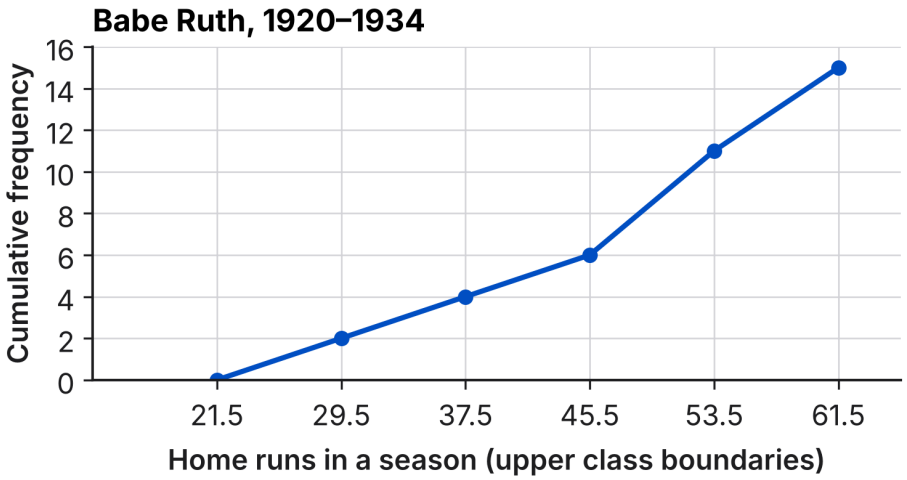
Example 3 Draw a frequency polygon of Ruth's home runs.



DEFINITION Ogive

An **ogive** (oh-jive) is a cumulative frequency graph. It plots the upper class boundaries against the cumulative frequencies, starting at the first lower boundary with a frequency of 0.

Example 4 Draw an ogive of Ruth's home runs. Then answer the question.



In how many of these seasons did Ruth hit 45 or fewer home runs? 6

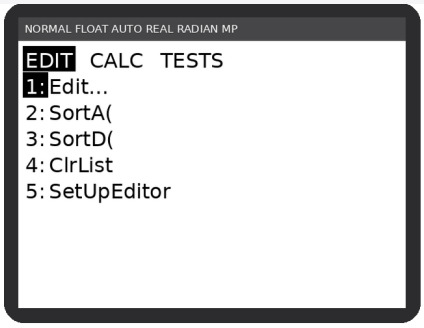
Read the ogive at 45.5: cumulative frequency 6.

TIP TI-84: Histogram

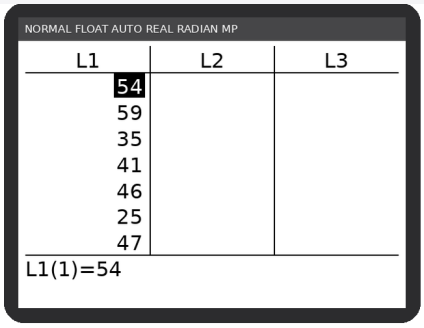
Enter the data: **STAT** ▸ **1:Edit**, type the data in **L1**.

Turn on the plot: **2nd** [STAT PLOT] ▸ **1:Plot1** ▸ **On**, Type: histogram (third picture), Xlist: **L1**, Freq: **1**.

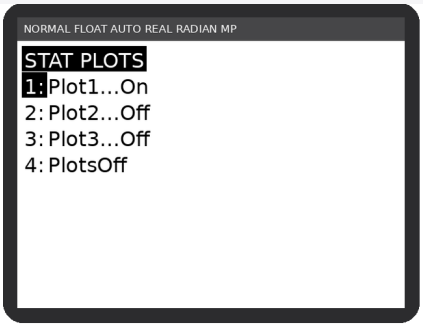
Set the classes: **WINDOW** ▸ Xmin = 21.5, Xmax = 61.5, Xscl = 8 (the class width), Ymin = 0, Ymax = 6. Press **GRAPH**, then **TRACE** to see each class and its count.



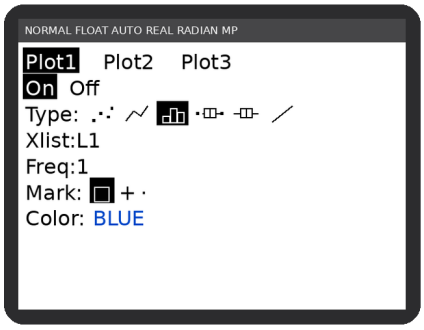
1 STAT ▸ 1:Edit



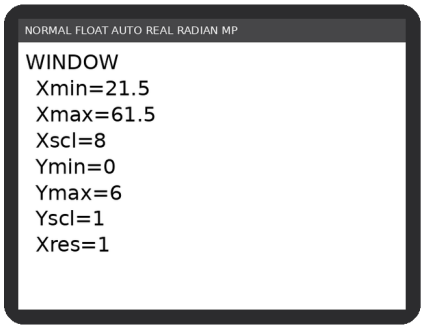
2 Type the data in L1



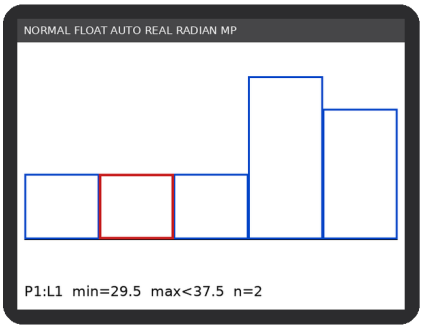
3 2nd Y= (STAT PLOT) ▸ 1:Plot1



4 On, histogram (third picture), Xlist: L1, Freq: 1



5 WINDOW: class boundaries, Xscl = class width



6 GRAPH, then TRACE: each bar's class and count

CHECK YOUR UNDERSTANDING

A frequency distribution has classes 10–19, 20–29, 30–39, and 40–49. What is the class width?

(A) 9
(B) 10 20 − 10 = 10 (subtract lower limits)
(C) 19
(D) 40

TOPIC 2.2

More Graphs and Displays

OBJECTIVES Graph quantitative data with stem-and-leaf plots and dot plots. Graph qualitative data with pie charts and Pareto charts. Graph paired data with scatter plots and time series charts.

GETTING STARTED Roger Maris broke Babe Ruth's record with 61 home runs in 1961. Here are Maris's home run totals for 1957 to 1968 (Source: Baseball-Reference):

14, 28, 16, 39, 61, 33, 23, 26, 8, 13, 9, 5

Do you think 61 was a typical season for Maris? Explain.

No. Most of his seasons were under 40. 61 is far above all of his other totals.

TIP
Hold onto this question. In 2.5 the class proves 61 is an outlier.

Graphing Quantitative Data

DEFINITION Stem-and-Leaf Plot

In a **stem-and-leaf plot**, each number is split into a **stem** (the leftmost digits) and a **leaf** (the rightmost digit). It shows the shape of the data and keeps every original data entry. Always include a **key**, such as 6 | 1 = 61. Leaves are written in order, least to greatest.

Example 1 Make a stem-and-leaf plot of Maris's home runs.

Stem	Leaves
0	5 8 9
1	3 4 6
2	3 6 8
3	3 9
4	
5	
6	1

Key: 6 | 1 = 61 home runs

DEFINITION **Back-to-Back Stem-and-Leaf Plot**

A **back-to-back** stem-and-leaf plot compares two data sets that share the same stems. Leaves for one set go to the left, read outward from the stem.

Example 2 The back-to-back plot compares Ruth (1920–1934) and Maris (1957–1968).

Ruth	Stem	Maris
	0	5 8 9
	1	3 4 6
5 2	2	3 6 8
5 4	3	3 9
9 7 6 6 6 1 1	4	
9 4 4	5	
0	6	1

Key: 5 | 4 | = 45 for Ruth, | 1 | 3 = 13 for Maris

Watch out: check the stems carefully. In this plot the stems are tens of home runs.

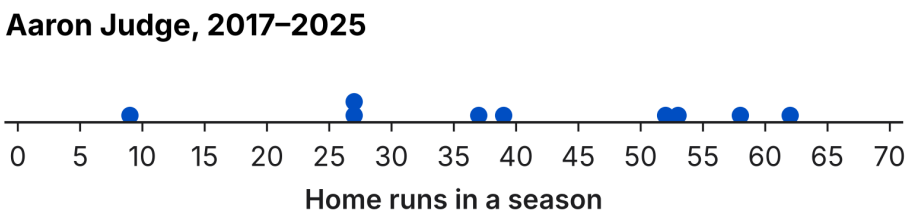
a.) Compare the two players. Who usually hit more home runs? Ruth

Most of Ruth's leaves sit on stems 4 and 5.
Most of Maris's leaves sit on stems 0, 1, and 2.

DEFINITION **Dot Plot**

In a **dot plot**, each entry is plotted as a point above a number line. Repeated values stack up.

Example 3 Make a dot plot of Aaron Judge's home runs for 2017 to 2025: 52, 27, 27, 9, 39, 62, 37, 58, 53. (Source: Baseball-Reference)



What might explain the lone 9? the 2020 season was only 60 games

2020 was a 60-game season, and Judge played only 28 games.

Graphing Qualitative Data

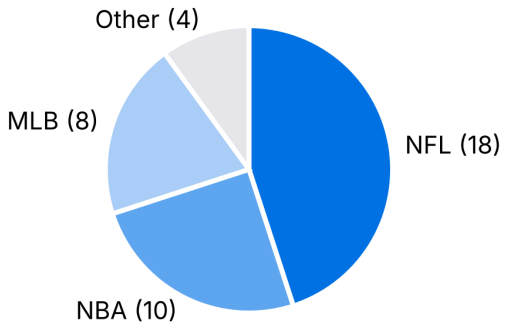
DEFINITION Pie Chart and Pareto Chart

A **pie chart** is a circle divided into sectors that represent categories. Each sector's angle is the relative frequency times 360° .

A **Pareto chart** is a vertical bar graph with the bars in order of decreasing frequency . It shows the most common categories first.

Example 4 A class of 40 students named their favorite pro league.

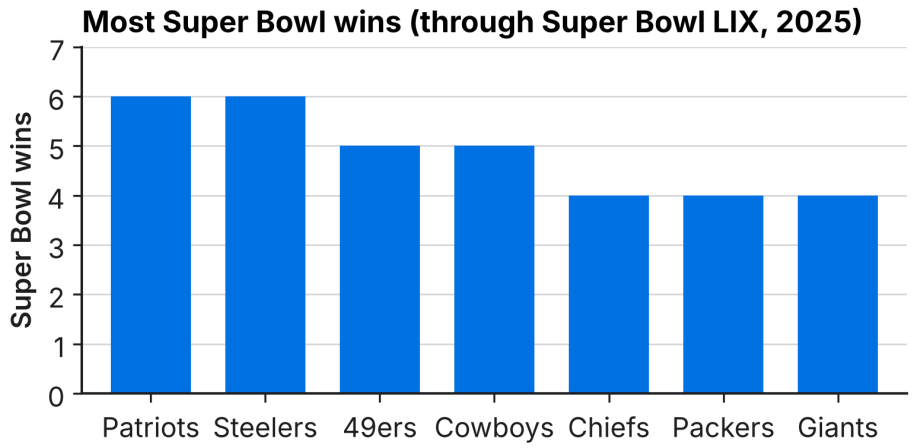
Favorite league: a class of 40 students



League	Frequency	Relative frequency	Angle
NFL	18	0.45	162°
NBA	10	0.25	90°
MLB	8	0.20	72°
Other	4	0.10	36°

NFL: $\frac{18}{40} = 0.45$ angle $= 0.45(360°) = 162°$

Example 5 The Pareto chart shows the NFL teams with the most Super Bowl wins.



a.) Which teams are tied for the most wins? Patriots and Steelers

b.) How many Super Bowls did these seven teams win in all? 34

$6+6+5+5+4+4+4=34$

Graphing Paired Data

DEFINITION **Scatter Plot and Time Series Chart**

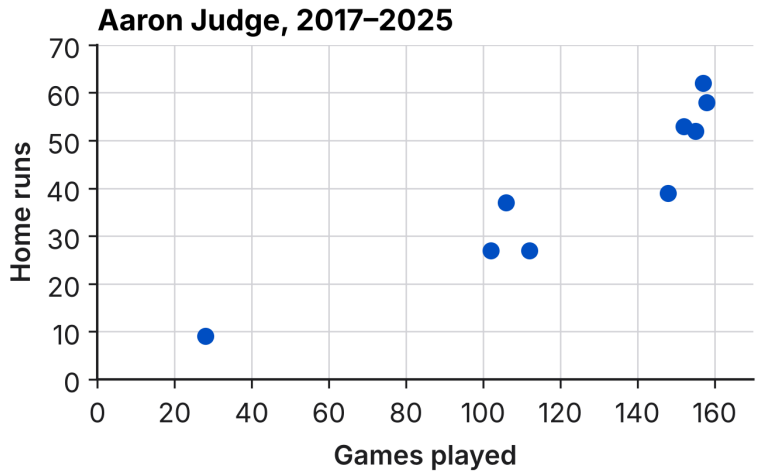
Paired data match each entry in one data set with an entry in a second data set, like (games, home runs).

A **scatter plot** graphs paired data as ordered pairs (x, y). It can show a relationship between two variables.

A **time series chart** graphs data entries that are measured over time, with time on the horizontal axis.

Example 6 Make a scatter plot of Judge's games played and home runs. Then describe the relationship.

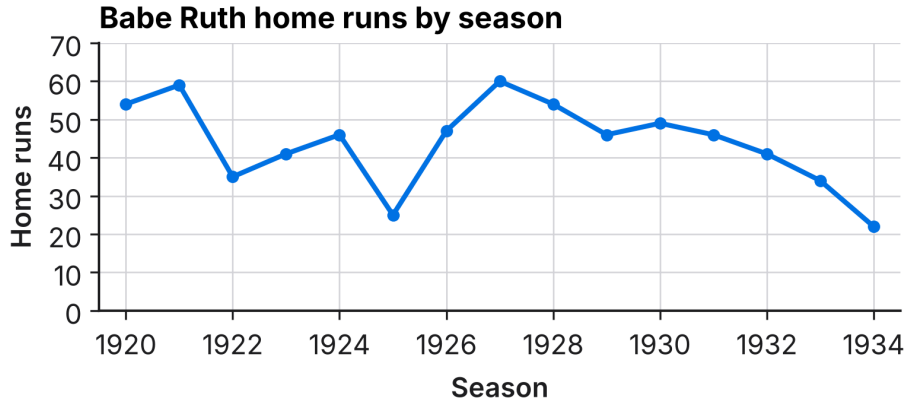
Season	2017	2018	2019	2020	2021	2022	2023	2024	2025
Games	155	112	102	28	148	157	106	158	152
HR	52	27	27	9	39	62	37	58	53



Relationship: more games, more home runs

As games played increase, home runs tend to increase.

Example 7 The time series chart shows Ruth's home runs by season.



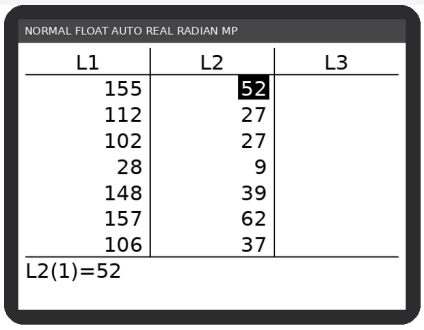
a.) In which season did Ruth hit the most home runs? 1927 (60)

b.) Describe the trend after 1930. steadily decreasing

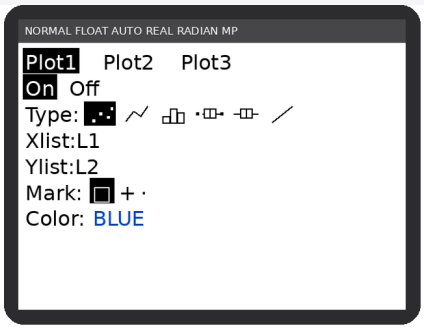
49, 46, 41, 34, 22: each season lower than the one before.

TIP TI-84: Scatter Plot

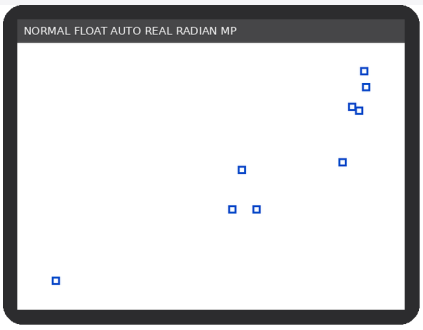
Enter x values in **L1** and y values in **L2**. **2nd** [STAT PLOT] ▸ **Plot1** ▸ **On**, Type: scatter (first picture), Xlist: **L1**, Ylist: **L2**. Press **ZOOM** ▸ **9:ZoomStat**.



❶ Games in L1, home runs in L2



❷ Plot1: On, scatter (first picture), Xlist: L1, Ylist: L2



❸ ZOOM ▸ 9:ZoomStat

CHECK YOUR UNDERSTANDING

Which graph is best for showing how one player's batting average changed from season to season?

(A) pie chart
(B) Pareto chart
(C) time series chart
(D) stem-and-leaf plot

data measured over time

TOPIC 2.3

Measures of Central Tendency

OBJECTIVES Find the mean, median, and mode of a data set. Find a weighted mean and the mean of a frequency distribution. Describe the shape of a distribution.

GETTING STARTED Maris hit 61 home runs in 1961, but under 40 in every other season. If you describe a "typical" Maris season with one number, should the 61 count? Explain.

It is part of the data, so it counts. But it is so unusual that it could pull an average up and make a typical season look bigger.

TIP
No right answer yet. This sets up mean vs. median and outliers.

Mean, Median, and Mode

DEFINITION Measures of Central Tendency

A **measure of central tendency** is a value that represents a typical, or central, entry of a data set.

Mean: the sum of the entries divided by the number of entries. Population mean $\mu = \frac{\Sigma x}{N}$.

Sample mean $\bar{x} = \frac{\Sigma x}{n}$.

Median: the middle entry when the data are in order. With an even number of entries, it is the mean of the two middle entries.

Mode: the entry that occurs most often. A data set can have one mode, more than one mode, or no mode. It is the only measure that works for qualitative data.

Example 1 Find the mean, median, and mode of Ruth's home runs, 1920–1934.

22, 25, 34, 35, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60

a.) Mean about 43.9

$\bar{x} = \frac{\Sigma x}{n}$

$\bar{x} = \frac{659}{15} \approx 43.9$

b.) Median 46

15 entries: the middle entry is the 8th.

median = 46

c.) Mode 46

46 occurs 3 times, more than any other entry.

Example 2 Find the mean, median, and mode of Maris's home runs, 1957–1968.

5, 8, 9, 13, 14, 16, 23, 26, 28, 33, 39, 61

a.) Mean about 22.9

$\bar{x} = \frac{\Sigma x}{n}$

$\bar{x} = \frac{275}{12} \approx 22.9$

b.) Median 19.5

12 entries: average the 6th and 7th.

$\text{median} = \frac{16 + 23}{2} = 19.5$

c.) Mode none

No entry repeats, so there is no mode.

Outliers

DEFINITION Outlier

An **outlier** is a data entry that is far removed from the other entries. An outlier can strongly affect the mean , but it has little effect on the median .

Example 3 Suppose Maris had hit 31 home runs in 1961 instead of 61.

a.) Find the new mean. about 20.4

$\Sigma x = 275 - 61 + 31 = 245$

$\bar{x} = \frac{245}{12} \approx 20.4$

b.) Find the new median. 19.5

ordered: 5, 8, 9, 13, 14, 16, 23, 26, 28, 31, 33, 39

median $= \frac{16 + 23}{2} = 19.5$ (no change)

c.) Which measure better describes a typical Maris season? Explain.

The median. It is resistant to the outlier 61.
The mean is pulled up by that one season.

TIP
Ask: if a billionaire walks into the room, what happens to the mean income? To the median?

TIP Choosing a Measure

Use the **mean** when the data have no outliers: it uses every entry. Use the **median** when the data have outliers or are skewed. Use the **mode** for qualitative data, such as a favorite team.

Weighted Mean and Grouped Data

DEFINITION Weighted Mean

A **weighted mean** is the mean of entries that have different weights w: $\bar{x} = \frac{\Sigma(x \cdot w)}{\Sigma w}$.

Example 4 A course grade is 50% tests, 20% homework, and 30% final exam. A student has a test average of 86, a homework average of 95, and a final exam score of 78. Find the course grade. 85.4

$$\bar{x} = \frac{\Sigma(x \cdot w)}{\Sigma w}$$

$$\bar{x} = \frac{86(0.50) + 95(0.20) + 78(0.30)}{0.50 + 0.20 + 0.30}$$

$$\bar{x} = \frac{43 + 19 + 23.4}{1} = 85.4$$

DEFINITION Mean of a Frequency Distribution

For grouped data, use each class midpoint x and frequency f: $\bar{x} = \frac{\Sigma(x \cdot f)}{n}$, where $n = \Sigma f$.

Example 5 Use the frequency distribution of Ruth's home runs from 2.1 to estimate his mean.

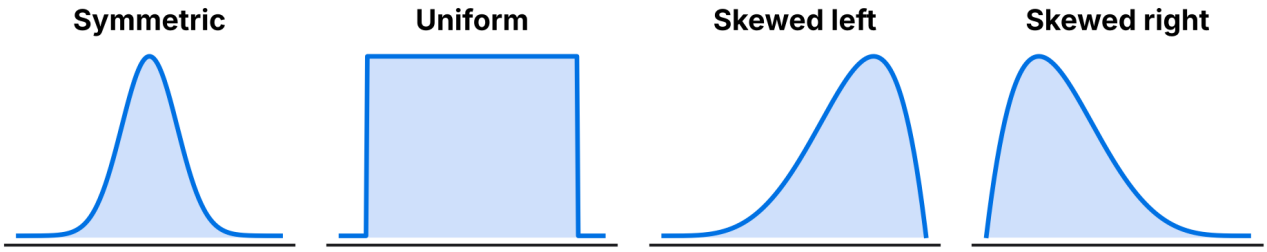
Midpoint, x	Frequency, f	x · f
25.5	2	51
33.5	2	67
41.5	2	83
49.5	5	247.5
57.5	4	230
	n = 15	Σ = 678.5

Estimated mean about 45.2

$$\bar{x} = \frac{\Sigma(x \cdot f)}{n} = \frac{678.5}{15} \approx 45.2$$

(The actual mean is 43.9. Grouping gives an estimate.)

Shapes of Distributions



DEFINITION Shape and Center

Symmetric: the left and right halves are mirror images. Mean = median.

Uniform: all entries or classes have about equal frequencies.

Skewed left: the tail extends to the left. The mean is usually less than the median.

Skewed right: the tail extends to the right. The mean is usually greater than the median.

TIP Remember

The mean gets pulled toward the **tail**. The tail tells the direction of the skew.

Example 6 Use the means and medians from Examples 1 and 2 to describe each shape.

a.) Ruth: mean 43.9, median 46 skewed left

mean < median: a tail of low seasons (22, 25)
pulls the mean down.

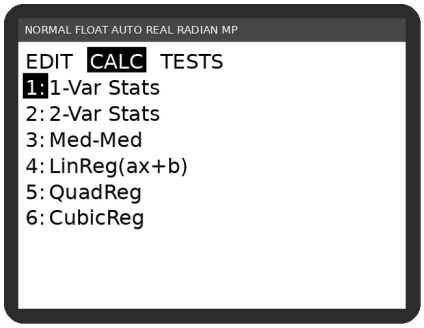
b.) Maris: mean 22.9, median 19.5 skewed right

mean > median: the 61 stretches the tail right
and pulls the mean up.

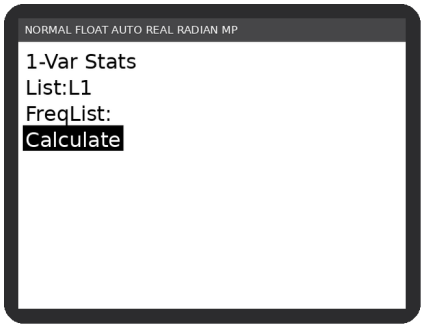
TIP TI-84: 1-Var Stats

Enter the data in **L1**. Press **STAT** ▸ **CALC** ▸ **1:1-Var Stats**, List: **L1**, then **Calculate**. The screen shows $\bar{x}$, Σx, n, and (scroll down) Med, minX, and maxX.

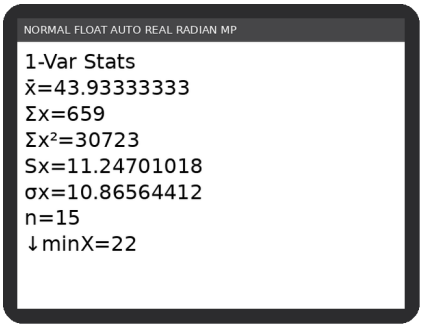
For a frequency distribution, put midpoints in **L1** and frequencies in **L2**, and set FreqList: **L2**.



❶ STAT ▸ CALC ▸ 1:1-Var Stats



❷ List: L1, leave FreqList blank, Calculate



❸ x̄, Σx, n; scroll down for Med, minX, maxX

CHECK YOUR UNDERSTANDING

A data set has a mean of 52 and a median of 45. What is the most likely shape of its distribution?

(A) symmetric
(B) uniform
(C) skewed left
(D) skewed right

mean > median: pulled right by the tail

TOPIC 2.4

Measures of Variation

OBJECTIVES Find the range, variance, and standard deviation of a data set. Interpret standard deviation. Use the Empirical Rule and Chebyshev's Theorem. Find the coefficient of variation.

GETTING STARTED Two hitters each average 30 home runs over three seasons. Hitter A hits 29, 30, and 31. Hitter B hits 10, 30, and 50. The means are the same. How are the hitters different?

Hitter A is consistent: every season is close to 30.
Hitter B is spread out: his seasons are far from 30.
The center is the same, but the spread is not.



Range

DEFINITION Range

The **range** of a quantitative data set is the difference between the maximum and minimum entries: Range = (Maximum entry) – (Minimum entry).

Example 1 Find the range of each data set.

a.) Maris, 1960–1963: 39, 61, 33, 23 38

Range = 61 − 23 = 38

b.) Ruth, 1920–1934 (from 2.3): min 22, max 60 38

Range = 60 − 22 = 38

c.) The ranges are equal. Does that mean the two data sets have the same spread? Explain.

No. The range uses only the two extreme entries.
It ignores how the other entries are spread out.

Variance and Standard Deviation

DEFINITION Deviation, Variance, and Standard Deviation

The **deviation** of an entry x is its distance from the mean: $x - \mu$ for a population, $x - \bar{x}$ for a sample. The sum of the deviations is always 0 .

Population: variance $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$ and standard deviation $\sigma = \sqrt{\sigma^2}$.

Sample: variance $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$ and standard deviation $s = \sqrt{s^2}$.

The **standard deviation** measures the typical distance of an entry from the mean. The more spread out the data, the greater the standard deviation.

1. **PROCEDURE Finding a Standard Deviation by Hand**

1. Find the mean.
2. Subtract the mean from each entry to get the deviations.
3. Square each deviation.
4. Add the squares to get the sum of squares, SS.
5. Divide by N (population) or $n - 1$ (sample) to get the variance.
6. Take the square root.

Example 2 Roger Maris hit 39, 61, 33, and 23 home runs from 1960 to 1963. Treat these 4 seasons as a population.

x	x – μ	(x – μ)²
39	0	0
61	22	484
33	-6	36
23	-16	256
Σx = 156	Σ = 0	SS = 776

a.) Find the mean. 39

$\mu = \frac{\Sigma x}{N} = \frac{156}{4} = 39$

b.) Find the population variance. 194

$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} = \frac{776}{4} = 194$

c.) Find the population standard deviation. about 13.9

$\sigma = \sqrt{194} \approx 13.9$ home runs

Example 3 Now treat the same 4 seasons as a sample of Maris's career.

a.) Find the sample variance. about 258.7

$s^2 = \frac{SS}{n-1} = \frac{776}{3} \approx 258.7$

b.) Find the sample standard deviation. about 16.1

$s = \sqrt{258.7} \approx 16.1$ home runs

c.) Why is s larger than σ?

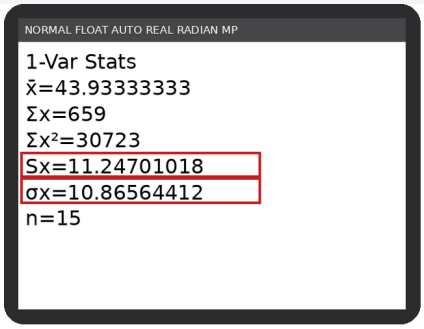
Dividing by $n - 1 = 3$ instead of 4 gives a larger result. This corrects for a sample usually being less spread out than its population.

TIP
Students only need: sample → $n - 1$, population → N. The "why" can stay light.

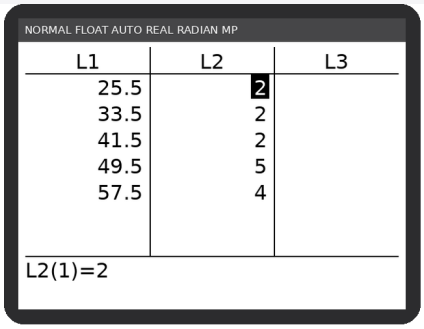
TIP TI-84: Standard Deviation

Enter the data in **L1**. Press **STAT** ▸ **CALC** ▸ **1:1-Var Stats**, then **Calculate**. **Sx** is the sample standard deviation s. **σx** is the population standard deviation σ. Square either one to get the variance.

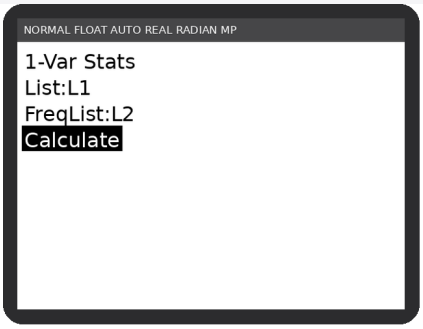
For a frequency distribution, put midpoints in **L1**, frequencies in **L2**, and set FreqList: **L2**.



1 Sx = sample s, σx = population σ



2 Frequency distribution: midpoints in L1, frequencies in L2



3 Set FreqList: L2, then Calculate

Example 4 Use a TI-84 and 1-Var Stats on the career data from 2.3. Treat each as a sample.

a.) Ruth, 1920–1934: $\bar{x} \approx 43.9$, $s \approx$ 11.2

b.) Maris, 1957–1968: $\bar{x} \approx 22.9$, $s \approx$ 16.0

c.) Which player's home run totals were more spread out? Explain.

Maris. His standard deviation is larger, so his seasons were farther from his mean, on average.

Interpreting Standard Deviation

DEFINITION **Coefficient of Variation**

The **coefficient of variation** (CV) compares the standard deviation to the mean:

$CV = \frac{s}{\bar{x}} \cdot 100\%$

Use it to compare the variation of data sets with very different means. A larger CV means more relative variation.

Example 5 Use Example 4 to find each player's coefficient of variation.

a.) Ruth about 25.6%

$CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{11.25}{43.93} \cdot 100\% \approx 25.6\%$

b.) Maris about 69.7%

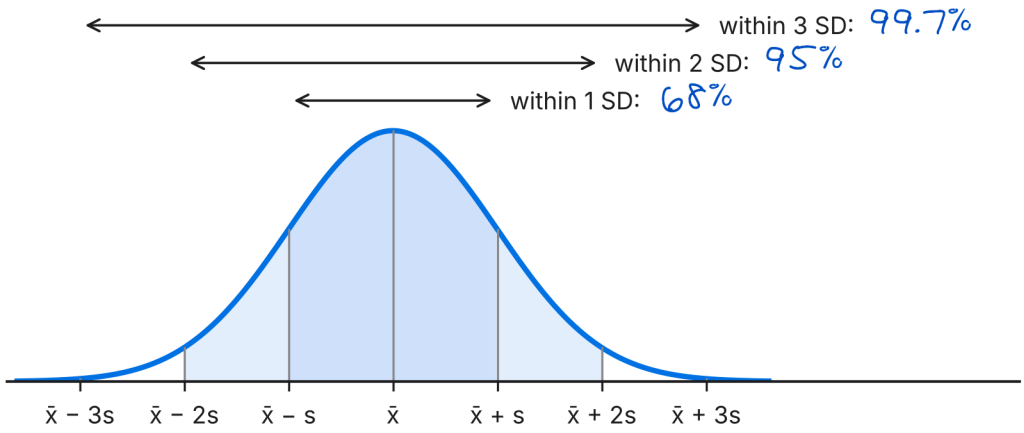
$CV = \frac{15.98}{22.92} \cdot 100\% \approx 69.7\%$

The Empirical Rule

DEFINITION **Empirical Rule (68–95–99.7 Rule)**

For data with a symmetric, **bell-shaped** distribution:

About 68% of the data lie within 1 standard deviation of the mean. About 95% lie within 2 standard deviations. About 99.7% lie within 3 standard deviations.



Example 6 Suppose the heights of a group of basketball players are bell-shaped with a mean of 79 inches and a standard deviation of 3.5 inches. Estimate the percent of players with each height.

a.) between 75.5 and 82.5 inches 68%

$79 - 3.5 = 75.5$ and $79 + 3.5 = 82.5$

within 1 SD: about 68%

b.) between 72 and 86 inches 95%

$79 - 2(3.5) = 72$ and $79 + 2(3.5) = 86$

within 2 SD: about 95%

c.) between 79 and 86 inches 47.5%

mean to 2 SD above is half of 95%:

$\frac{95\%}{2} = 47.5\%$

TIP
Have students sketch the curve and shade. Halving by symmetry is the key move.

d.) taller than 82.5 inches 16%

outside 1 SD: $100\% - 68\% = 32\%$

upper tail is half: $\frac{32\%}{2} = 16\%$

Chebyshev's Theorem

DEFINITION Chebyshev's Theorem

For **any** data set, no matter its shape, the portion of the data within k standard deviations of the mean ($k > 1$) is **at least** $1 - \frac{1}{k^2}$.

$k = 2$: at least 75% . $k = 3$: at least 88.9% .

Example 7 Maris's 1957–1968 seasons are skewed right, so the Empirical Rule does not apply. Use $\bar{x} \approx 22.9$ and $s \approx 16.0$.

a.) Find the interval within 2 standard deviations of the mean. -9.1 to 54.9

$22.9 - 2(16.0) = -9.1$

$22.9 + 2(16.0) = 54.9$

b.) By Chebyshev's Theorem, at least what percent of his seasons lie in this interval? 75%

$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = 0.75 = 75\%$

c.) What percent actually do? about 91.7%

All but the 61 lie in the interval:

$\frac{11}{12} \approx 91.7\%$ (at least 75%: true)

CHECK YOUR UNDERSTANDING

A data set is bell-shaped with a mean of 50 and a standard deviation of 5. About what percent of the data lie between 40 and 60?

(A) 68%
(B) 75%
(C) 95% 40 and 60 are 2 SD from 50: about 95%
(D) 99.7%

TOPIC 2.5

Measures of Position

OBJECTIVES Find the quartiles and interquartile range of a data set. Draw a box-and-whisker plot. Find percentiles. Find and interpret z-scores.

GETTING STARTED Babe Ruth hit 60 home runs in 1927. Roger Maris hit 61 in 1961. Which season stood out more compared with the rest of that player's own career? Make a guess.

Guess: Maris. His 61 was far above his other seasons, while Ruth hit 50+ many times.

TIP
Take a vote and save it. Example 7 settles it with z-scores.

Quartiles

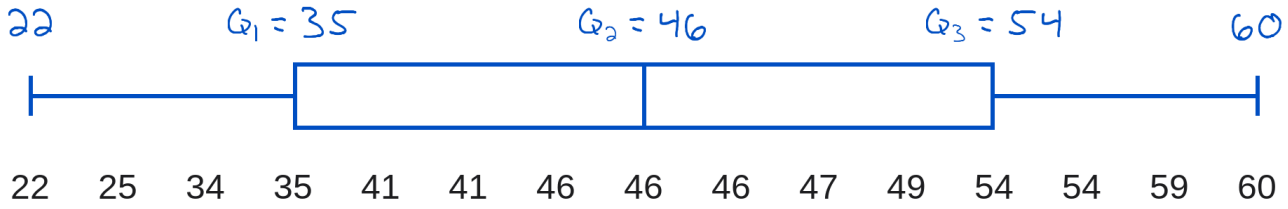
DEFINITION Quartiles

The three **quartiles** Q_1, Q_2, and Q_3 divide an ordered data set into four parts with about the same number of entries.

Q_2 is the median. Q_1 is the median of the entries below Q_2. Q_3 is the median of the entries above Q_2.

The **interquartile range** is IQR = $Q_3 - Q_1$. It measures the spread of the middle 50% of the data.

Example 1 Find the quartiles and IQR of Ruth's home runs, 1920–1934.



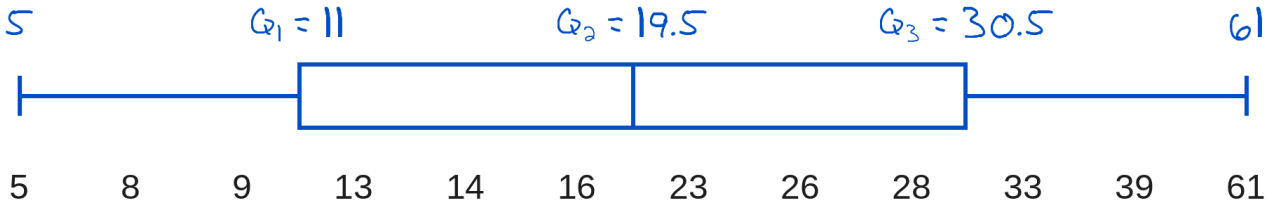
a.) Q_2 46 Q_1 35 Q_3 54

15 entries: Q_2 is the 8th entry = 46
lower 7: 22, 25, 34, 35, 41, 41, 46 → Q_1 = 35
upper 7: 46, 47, 49, 54, 54, 59, 60 → Q_3 = 54

b.) IQR 19

$IQR = Q_3 - Q_1 = 54 - 35 = 19$

Example 2 Find the quartiles and IQR of Maris's home runs, 1957–1968.



a.) Q_2 19.5 Q_1 11 Q_3 30.5

12 entries: $Q_2 = \frac{16+23}{2} = 19.5$

lower 6: 5, 8, 9, 13, 14, 16 → $Q_1 = \frac{9+13}{2} = 11$

upper 6: 23, 26, 28, 33, 39, 61 → $Q_3 = \frac{28+33}{2} = 30.5$

b.) IQR 19.5

$IQR = 30.5 - 11 = 19.5$

DEFINITION **Identifying Outliers with the IQR**

An entry is an **outlier** if it is less than $Q_1 - 1.5(\text{IQR})$ or greater than $Q_3 + 1.5(\text{IQR})$. These two values are called fences .

Example 3 Is Maris's 61 an outlier? Use the fences.

$1.5(IQR) = 1.5(19.5) = 29.25$

lower fence: $11 - 29.25 = -18.25$

upper fence: $30.5 + 29.25 = 59.75$

$61 > 59.75$, so 61 is an outlier.

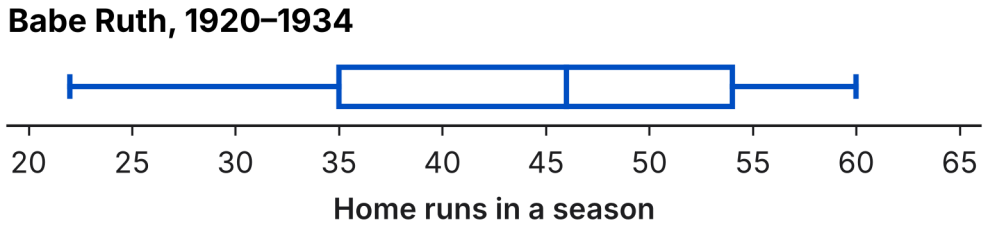
Box-and-Whisker Plots

DEFINITION Five-Number Summary and Box Plot

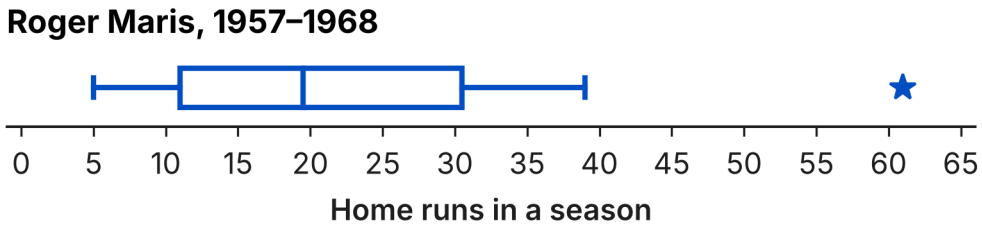
The **five-number summary** is the minimum, Q_1, Q_2, Q_3, and maximum. A **box-and-whisker plot** (box plot) displays it: a box from Q_1 to Q_3, a vertical line at the median, and whiskers out to the minimum and maximum.

In a **modified** box plot, outliers are marked with a star, and the whisker stops at the most extreme entry that is not an outlier.

Example 4 Write the five-number summary for Ruth: 22, 35, 46, 54, 60. Then draw the box plot.



Example 5 Draw a modified box plot for Maris. Mark the outlier with a star.



Compare the two plots. What do they tell you?

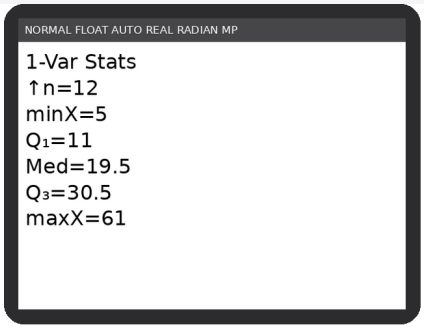
Ruth's box sits much farther right: higher totals.
Ruth: left whisker longer (skewed left).
Maris: right side longer, plus an outlier (skewed right).

TIP
Have students point to the median line, not the middle of the box. They often mix these up.

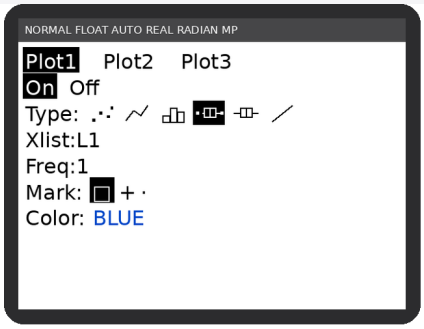
TIP TI-84: Five-Number Summary and Box Plot

Enter the data in **L1**. **1-Var Stats** shows minX, Q_1, Med, Q_3, and maxX (scroll down).

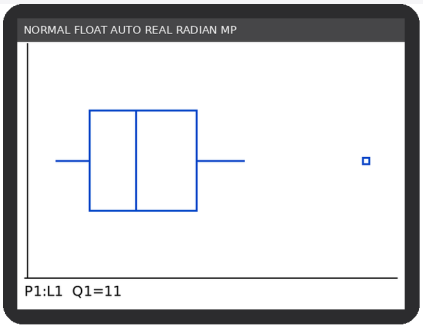
Press **2nd Y=** (STAT PLOT) ▸ **1:** Plot1 ▸ **On**. Choose the first box plot icon (modified, shows outliers). Xlist: **L1**. Press **ZOOM 9** (ZoomStat).



❶ 1-Var Stats, scroll down: minX, Q₁, Med, Q₃, maxX



❷ Plot1: On, first box plot icon (modified), Xlist: L1



❸ ZOOM 9 (ZoomStat). The outlier shows as a small square.

Percentiles

DEFINITION Percentiles

The **percentiles** divide an ordered data set into 100 parts. If an entry is at the 80th percentile, then about 80% of the entries are at or below it.

Percentile of $x = \dfrac{\text{number of entries less than } x}{\text{total number of entries}} \cdot 100$, rounded to the nearest whole number.

Q_1, Q_2, and Q_3 are about the 25th, 50th, and 75th percentiles.

Example 6 Ruth hit 54 home runs in 1920 and in 1928. Find the percentile of 54 in his 1920–1934 data. 73rd

entries less than 54: 22, 25, 34, 35, 41, 41, 46, 46, 46, 47, 49

that is 11 entries

percentile = $\frac{11}{15} \cdot 100 \approx 73$ → 73rd percentile

The Standard Score (z-Score)

DEFINITION z-Score

The **standard score**, or **z-score**, tells how many standard deviations an entry x is from the mean: $z = \frac{x-\mu}{\sigma}$ (or use $\bar{x}$ and s for a sample).

A negative z-score means the entry is below the mean. A z-score less than –2 or greater than 2 is **unusual**. Less than –3 or greater than 3 is **very unusual**.

Example 7 Compare Ruth's 60 home runs in 1927 with Maris's 61 in 1961, each against his own career. Use the sample statistics.

a.) Ruth: $\bar{x} \approx 43.93$, $s \approx 11.25$ $z \approx 1.43$

$$z = \frac{x - \bar{x}}{s} = \frac{60 - 43.93}{11.25} \approx 1.43$$

b.) Maris: $\bar{x} \approx 22.92$, $s \approx 15.98$ $z \approx 2.38$

$$z = \frac{61 - 22.92}{15.98} \approx 2.38$$

c.) Which season was more unusual for that player? Explain.

Maris's 61. It is 2.38 SDs above his mean (unusual, since $z > 2$). Ruth's 60 is only 1.43 SDs above his.

TIP
Point out: 61 vs. 60 looks like a tie. z-scores compare each to its own data set. Check the vote from Getting Started.

CHECK YOUR UNDERSTANDING

A test has a mean of 72 and a standard deviation of 8. What is the z-score of a test score of 60?

(A) –1.5 $z = (60 - 72)/8 = -1.5$

(B) –0.67

(C) 1.5
(D) 12