**AP Stats**
Unit 01 – Univariate Data
Day 2 Notes (1.1-1.2)

Name Key

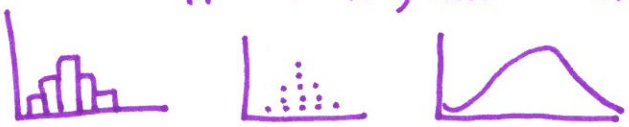
**OVERALL PATTERN OF A DISTRIBUTION**

Distribution: of a variable tells us what values the variable takes and how often it takes those values.

**When describing all distributions, you must include the following:**

SHAPE: the main features: peaks, clusters, gaps, symmetry/skewed distributions

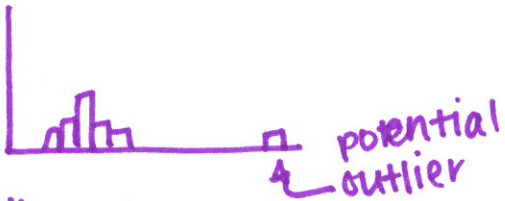
a) Symmetric: if the right and left sides of the graph are approximately mirror images.



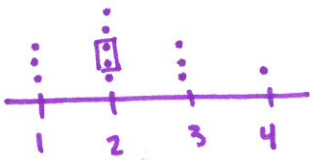
b) Skewed right: if the right side of the graph is much longer than the left side (long right tail).

c) Skewed left: if the left side of the graph is much longer than the right side (long left tail).

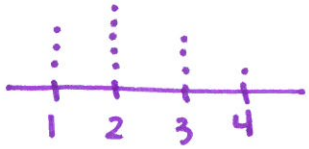
OUTLIERS: value(s) that differ somewhat from the overall pattern.




CENTER: the value or values that divide the observation in half
- the median or where the peak of the data is (most of the time)




SPREAD: tells us how much variability there is in the data
- the range of the data (largest # − smallest #)



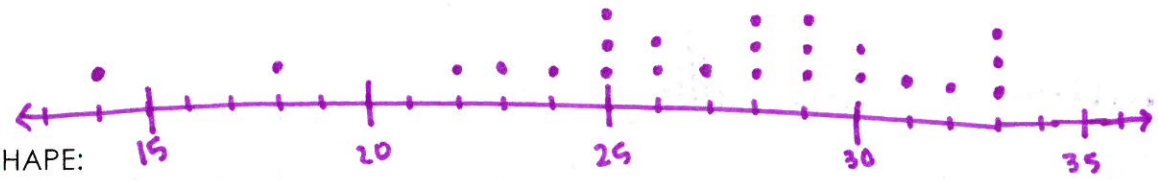

4 − 1 = 3
↑ range

**Example: Are You Driving a Gas Guzzler?**

The Environmental Protection Agency (EPA) is in charge of determining and reporting fuel economy rating for cars (think of those large window stickers on a new car). For years, consumers complained that their actual gas mileages were noticeably lower than all the values reported by the EPA. It seems that the EPA's tests – all of which are done on computerized devices to ensure consistency – did not consider things like outdoor temperature, use of the air conditioner, or realistic acceleration and braking by drivers. In 2008, the EPA changed the method for measuring a vehicle's fuel economy to try to give more accurate estimates. The table below displays the EPA estimates of highway gas mileage in miles per gallon (mpg) for a sample of 24 model year 2009 midsize cars.

| Model | MPG | Model | MPG | Model | MPG |
|---|---|---|---|---|---|
| Acura RL | 22 | Dodge Avenger | 30 | Mercury Milan | 29 |
| Audi A6 Quattro | 23 | Hyundai Elantra | 33 | Mitsubishi Galant | 27 |
| Bentley Arnage | 14 | Jaguar XF | 25 | Nissan Maxima | 26 |
| BMW 528I | 28 | Kia Optima | 32 | Rolls Royce Phantom | 18 |
| Buick Lacrosse | 28 | Lexus GS 350 | 26 | Saturn Aura | 33 |
| Cadillac CTS | 25 | Lincoln MKZ | 28 | Toyota Camry | 31 |
| Chevrolet Malibu | 33 | Mazda 6 | 29 | Volkswagen Passat | 29 |
| Chrysler Sebring | 30 | Mercedes-Benz E350 | 24 | Volvo S80 | 25 |

Create a dot plot with the given information and then describe the distribution.




SHAPE:

skewed left
peaks @ 25, 28-29, 33
gaps between 14-18, 18-22

OUTLIERS:

potential outliers at 14 and 18

CENTER:

median at 28

SPREAD:

range = 33-14 = 19

**HISTOGRAMS**

- used for quantitative variables
- make bins (3-7 allowed, 5-7 preferable)

**Example: Born Outside the US**

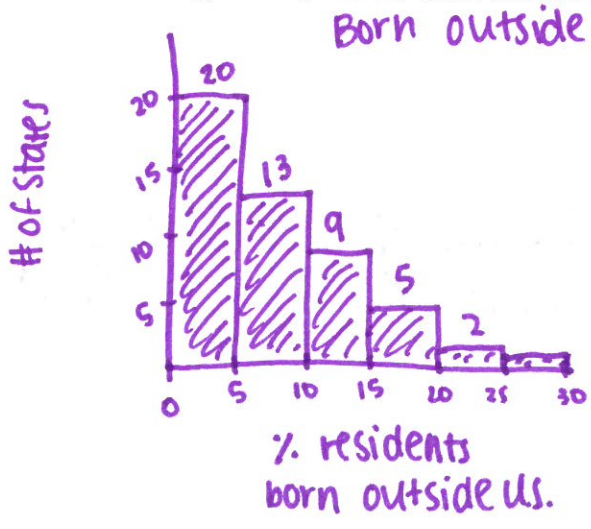
What percent of your home state's residents were born outside the United States? The country as a whole has 12.9% foreign-born residents, but the states vary from 1.2% in West Virginia to 27.2% in California. The table below presents the data for all 50 states. It is much easier to see from a graph than from the table how your state compares with other states.

| State | Percent | State | Percent | State | Percent |
|---|---|---|---|---|---|
| Alabama | 2.8 | Louisiana | 2.9 | Ohio | 3.6 |
| Alaska | 7.0 | Maine | 3.2 | Oklahoma | 4.9 |
| Arizona | 15.1 | Maryland | 12.2 | Oregon | 9.7 |
| Arkansas | 3.8 | Massachusetts | 14.1 | Pennsylvania | 5.1 |
| California | 27.2 | Michigan | 5.9 | Rhode Island | 12.6 |
| Colorado | 10.3 | Minnesota | 6.6 | South Caroline | 4.1 |
| Connecticut | 12.9 | Mississippi | 1.8 | South Dakota | 2.2 |
| Delaware | 8.1 | Missouri | 3.3 | Tennessee | 3.9 |
| Florida | 18.9 | Montana | 1.9 | Texas | 15.9 |
| Georgia | 9.2 | Nebraska | 5.6 | Utah | 8.3 |
| Hawaii | 16.3 | Nevada | 19.1 | Vermont | 3.9 |
| Idaho | 5.6 | New Hampshire | 5.4 | Virginia | 10.1 |
| Illinois | 13.8 | New Jersey | 20.1 | Washington | 12.4 |
| Indiana | 4.2 | New Mexico | 10.1 | West Virginia | 1.2 |
| Iowa | 3.8 | New York | 21.6 | Wisconsin | 4.4 |
| Kansas | 6.3 | North Carolina | 6.9 | Wyoming | 2.7 |
| Kentucky | 2.7 | North Dakota | 2.1 | | |

The *individuals* in this data set are the states.

The *variable* is the percent of a state's residents who are foreign-born.

Create a histogram by hand, then describe the distribution.




- tick mark included in upper bin (right)

**Percentile:** the value such that p % of the observations fall at (or below) it.

↑ allows for 100th %-ile

ex: 75% → you scored better than or equal to 75% of the people who took the test. You are in the top 25%!

Note: You cannot be in the 100th percentile. but you can be in the 0% percentile!

**Example:**
The test scores for 25 students are shown below.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 79 | 81 | 80 | 77 | 73 | 83 | 74 | 93 | 78 | 80 | 75 | 67 | 73 |
| 77 | 83 | 86 | 90 | 79 | 85 | 83 | 89 | 84 | 82 | 77 | 72 | |

1. Use the scores to find the percentiles for the following students

a) a score of 72

below 72: $Y_{25} = 0.04$ or **4th percentile**

b) a score of 93

below 93: $24/25 = 0.96$ or **96th percentile**

c) the 2 students who scored an 80

| S | L |
|---|---|
| 6 | 7 |
| 7 | 2 3 3 4 5 7 7 7 8 9 9 |
| 8 | 0 0 1 2 3 3 3 4 5 6 9 |
| 9 | 0 3 |

12 below them

$12/25 = 0.48$ or **48th percentile**

* you do not include your score!

**Cumulative Relative Frequency Graphs (O-GIVE)**

- uses percentiles (%-iles).
- uses Quantitative variables.
- describes the position of an individual within a distribution or is used to locate a specified percentile of the distribution.

**Example:**
Here is a frequency table that summarizes the ages of the first 44 US presidents when they were inaugurated:

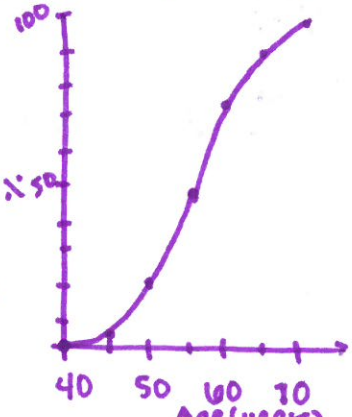
| Age | Frequency |
|---|---|
| 40-44 | 2 |
| 45-49 | 7 |
| 50-54 | 13 |
| 55-59 | 12 |
| 60-64 | 7 |
| 65-69 | 3 |

* we plot a point corresponding to the cumulative relative frequency in each class at the smallest value of the next class (like bins of a histogram)

Steps to create an O-GIVE:

1. Start with a frequency table.
2. Change to relative frequency: frequency ÷ total # of participants (to get a %).
3. find cumulative frequency: add the previous frequencies together to get the new frequency.
4. find the cumulative relative frequency: cumulative freq. ÷ total # of participants (to get a %).

| Age | freq. | rel. freq. | cumu. freq. | CRF |
|---|---|---|---|---|
| 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% |
| 55-59 | 12 | 12/44 = 27.3% | 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% |




In what percentile was Bill Clinton's age at inauguration? He was 46 when he took office.

~8th percentile

In what percentile was Ronald Reagan's age at inauguration? He was 69 when he took office.

99th percentile

What age at inauguration corresponds to the 50th percentile?

55 years

What is the range of ages at inauguration for the middle 50% of presidents?

52 - 59 years

**Time Plots**: where each observation is plotted against time (OVER TIME)

- time is plotted on the horizontal axis (x-axis) (or scale)
- variable of interest is plotted on the vertical axis (y-axis)
- connect lines between points so that we know it includes all points between the data points given.

Trend: a long-term upward or downward movement over time.

Seasonal variation: a pattern that repeats itself at regular time intervals.

ex: