AP Statistics Notes

Unit 1: Exploring One-Variable Data

Overview

You'll be introduced to how statisticians approach variation and practice representing data, describing distributions of data, and drawing conclusions based on a theoretical distribution.

1.1 What Can We Learn from Data?

Statistics is the science of collecting, organizing, analyzing, and interpreting data to make decisions or answer questions.

Key Terms

- **Individual:** A single unit from which data is collected (e.g., a person, object).
- **Variable:** A characteristic or measurement that varies from one individual to another.
- **Population:** The entire set of individuals we want to study.
- **Sample:** A subset of the population from which we actually collect data.

In statistics, we collect sample data on certain variables using measurements on individuals.

Example: A researcher surveys a random sample of 250 college students to examine their study habits. In this study, the population is all college students and the sample is the subset of 250 who took the survey. The variable here would be "hours studied per week" or some similar quantity of interest.

1.2 Variables and Types of Data

Key Terms

- **Categorical/Qualitative Variable:** A variable that can take on one of a limited, usually fixed number of possible "classes" or group labels.
- **Numerical/Quantitative Variable:** A variable that takes on number values, from a measured or counted quantity, and can be analyzed using arithmetic.

Numerical variables can be further split into either discrete or continuous variables.

Key Terms

- **Discrete Variable:** Have a countable set of possible number values
- **Continuous Variable:** Takes on an infinite range of possible values

Example:

- Ethnicity: Categorical
- Income: Quantitative (continuous)
- High school class: Categorical
- Age in years: Quantitative (discrete)

1.3–1.4 Representing Categorical Data with Tables and Graphs

Tables can be used to summarize data and get a quick overview of our data.

Key Terms

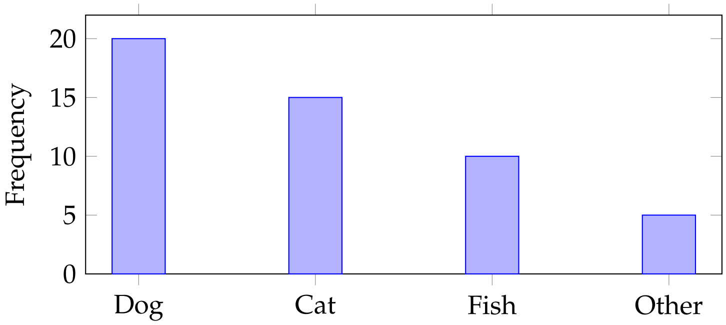
- The **frequency** of a value is just how many times that value occurs among our data.
- The **relative frequency** (rf) of a value is the ratio of its frequency to the total number of observations, n.
- The **cumulative frequency** (cf) of a value uses the number of observations less than or equal to that value. (This is useful when the possible values are *ordered* in some way.)
- We could similarly calculate the **relative cumulative frequency**.
- A **frequency table** gives the number of cases falling into each category. A **relative frequency table** includes the proportion of cases falling into each category.

Frequency Table

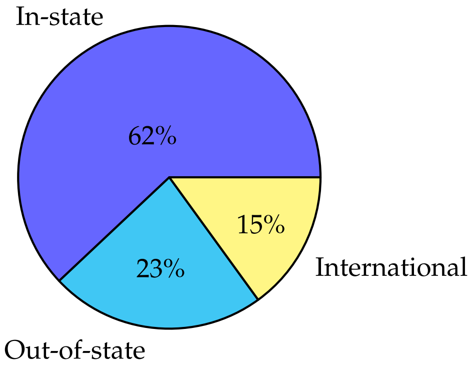
Pet	Frequency	Relative Frequency
Dog	20	0.40
Cat	15	0.30
Fish	10	0.20
Other	5	0.10

Of course, to get a better view of the data, we will want to use charts and graphs too. Here are some examples of graphs that are widely used for **data visualization**.

Bar Chart

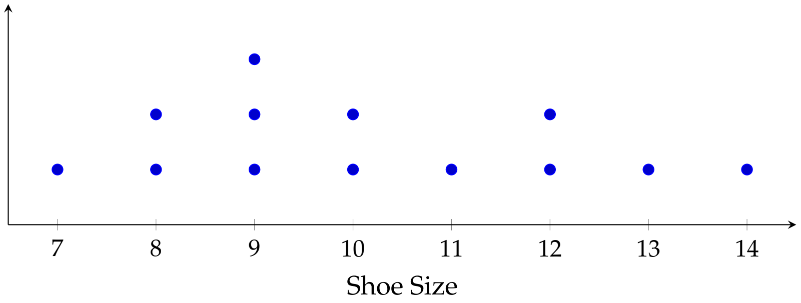


Pie Chart

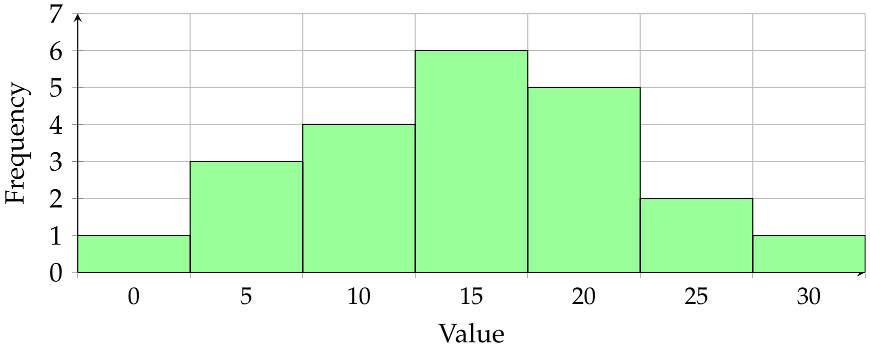


1.5 Representing Quantitative Data with Tables and Graphs

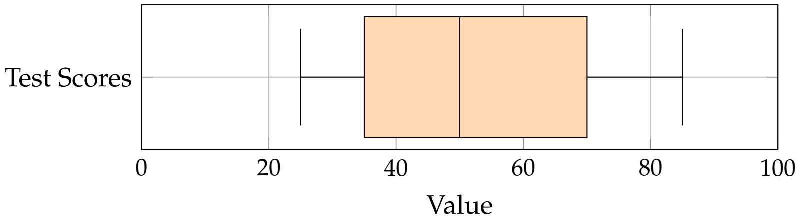
Dotplot



Histogram



Boxplot



Stemplot

Stem	Leaf
2	3
3	1 4 8
4	0 2 5
5	0 7
6	1

1.6 Describing the Distribution of Quantitative Data

When it comes to quantitative variables, we will want to discuss certain important characteristics of distributions.

Key Terms

SOCS:

Shape: The distribution could be **symmetric** or skewed in one direction, either **left-skewed** or **right-skewed** (see below example).

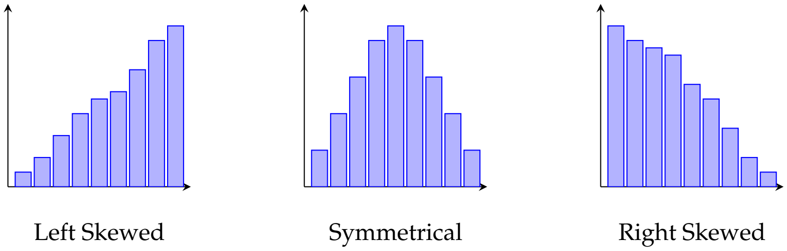
Outliers: Look for any highly unusual values that are far away from the bulk of the data.

Center: Estimate the typical or average value from the graph of the distribution.

Spread: The variability/dispersion of the data, perhaps give the highest and lowest values.

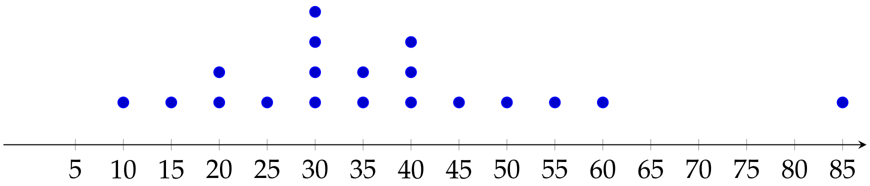
Other features: gaps, multiple peaks (unimodal vs. bimodal vs. multimodal), clusters, uniformity

Example:



Examples of Distribution Shapes

Practice: A dotplot shows the number of text messages sent by 15 students in one day. Describe the distribution using SOCS.



- Shape: Right-skewed
- Outliers: 85
- Center: About 30
- Spread: Data is quite spread out, ranges from 10 to 85

1.7 Summary Statistics

We've just seen how to briefly describe a distribution of data using its shape, outliers, center, and spread. Most likely though, we're going to want to gain a bit more information about and what exactly the data is telling us. That brings us to calculating **statistics** (the name of the course itself!).

We can think of the population as having the true characteristics like center and spread, and the sample is what we're using to approximate these characteristics with limited data. This is really the essence of statistics: using parts of the whole to study the bigger picture.

So this is *technically* what we mean when we refer to a statistic. Bottom line, finding statistics will let us describe distributions in much finer detail, particularly when talking about center, spread, and outliers.

Key Formulas

- **Mean**: $\bar{x} = \frac{1}{n} \sum x_i$
- **Mode**: The most commonly occurring value
- **Median**: Middle value
- **Range**: Maximum value - Minimum value
- **Quartiles**: Divide a data set into four equal parts
  - **Q1 (First Quartile)** is the median of the lower half of the data; it marks the 25th percentile
  - **Q3 (Third Quartile)** is the median of the upper half of the data; it marks the 75th percentile
- **Interquartile Range (IQR)**: $Q_3 - Q_1$
- **Five-Number Summary**: Minimum, Q1, Median, Q3, Maximum
- **Standard Deviation (SD)**: $s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$

Practice: Given the dataset = $\{2, 4, 4, 6, 8\}$, find the mean, median, and standard deviation.

Mean:

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{5}(2+4+4+6+8) = \frac{24}{5} = 4.8$$

Median = 4
SD:

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2} \\ &= \sqrt{\frac{1}{4}\sum(x_i - 4.8)^2} \\ &= \sqrt{\frac{1}{4}[(2-4.8)^2 + (4-4.8)^2 + (4-4.8)^2 + (6-4.8)^2 + (8-4.8)^2]} \\ &\approx 2.28 \end{aligned}$$

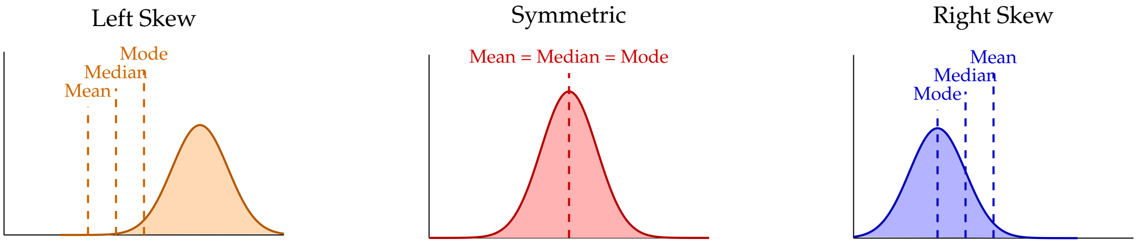
Practice: Find the five-number summary for a set of measurements:

$$\{2,\ 3,\ 5,\ 7,\ 8,\ 9,\ 10,\ 11,\ 12,\ 13,\ 15,\ 17,\ 18,\ 20,\ 22\}$$

Minimum = 2
$Q_1 = 7$ (Median of the lower half)
$Q_2 = 11$ (Median of the full dataset)
$Q_3 = 17$ (Median of the upper half)
Maximum = 22

$$\boxed{\text{Five-number summary: } [\text{Min.} = 2,\ Q_1 = 7,\ \text{Median} = 11,\ Q_3 = 17,\ \text{Max.} = 22]}$$

Depending on the shape of the distribution, we can say different things about its mean versus its median. In particular, see the following:



Let's talk about outliers for a bit. Specifically, the presence of outliers has a different effect on different summary statistics.

- The mean, standard deviation, and range are considered nonresistant (or non-robust) because they are influenced by outliers.
- The median and IQR are considered resistant (or robust), because outliers do not greatly (if at all) affect their value.

How do we detect whether a value is an outlier of the data set? There are several methods out there, but the most common is the **1.5 × IQR Criterion**. We'll demonstrate with an example.

Practice: Determine outliers using the 1.5 × IQR Criterion.

Dataset:

$$4,\ 5,\ 7,\ 8,\ 10,\ 11,\ 13,\ 15,\ 18,\ 19,\ 22,\ 25,\ 45$$

Step 1: Find the Quartiles There are 13 values, so the median (Q_2) is the 7th value:

$$Q_2 = 13$$

Lower half (below the median): $4,\ 5,\ 7,\ 8,\ 10,\ 11$ Q_1 is the median of this group:

$$Q_1 = \frac{7+8}{2} = 7.5$$

Upper half (above the median): $15,\ 18,\ 19,\ 22,\ 25,\ 45$ Q_3 is the median of this group:

$$Q_3 = \frac{19+22}{2} = 20.5$$

Step 2: Compute the IQR

$$\text{IQR} = Q_3 - Q_1 = 20.5 - 7.5 = 13$$

Step 3: Apply the 1.5 × IQR Rule

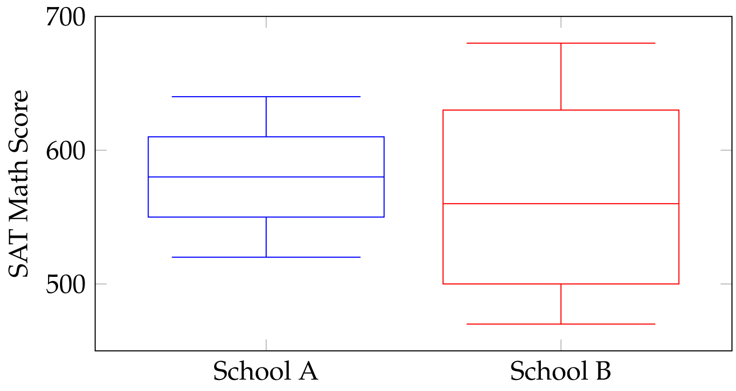
$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR} = 7.5 - 1.5 \times 13 = -12$$
$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR} = 20.5 + 1.5 \times 13 = 40$$

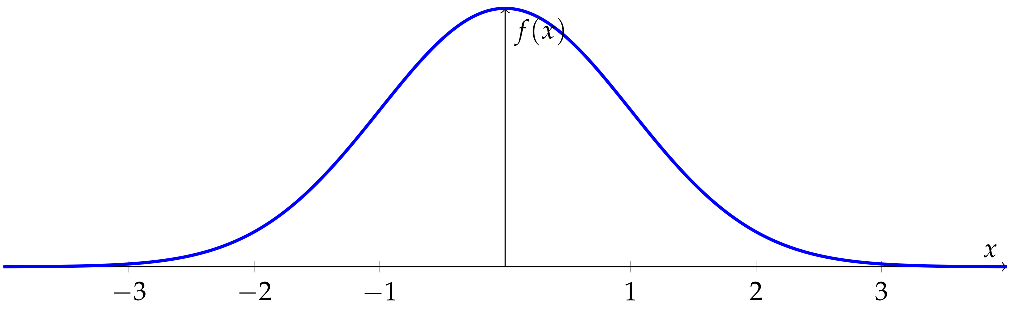
Step 4: Identify Outliers Any values less than the lower bound of -12 or greater than the upper bound of 40 are outliers. Only one value exceeds 40:

$$\boxed{\text{45 is an outlier.}}$$

1.8–1.9 Comparing Distributions

Practice: Compare the following boxplots of SAT math scores.





Normal Distribution

- **Shape and Skewness:** School A's boxplot is more symmetric, while School B shows a slight positive skew, with a longer whisker on the upper end. This may indicate more high outlier scores at School B or a few students scoring exceptionally well.
- **Outliers:** Neither School A nor School B seem to have any outlier values.
- **Center:** School A has a higher median SAT Math score (580) compared to School B (560), suggesting that typical scores are slightly higher at School A.
- **Spread:** School B has a wider interquartile range (IQR = 630 - 500 = 130) than School A (IQR = 610 - 550 = 60), indicating that the middle 50% of scores at School B are more spread out. School B also has a larger overall range (680 - 470 = 210) compared to School A (640 - 520 = 120), meaning the entire distribution of scores is more variable in School B.
- **Conclusion:** School A generally has higher typical scores and less variability, while School B has more spread in both the middle 50% and overall scores. Depending on the goal (e.g., high achievers vs. consistency), each school has its strengths.

1.10 The Normal Distribution

Key Formulas

Z-score: $z = \frac{x-\mu}{\sigma}$, where x is the observation, μ is the mean of the data, and σ is the SD of the data.

The Z-score, also sometimes called the **standardized score**, represents the number of standard deviations (SDs) away we are from the mean.

Example: Test score 85, scores have a mean of $\mu = 80$ and SD of $\sigma = 5$.

$$\boxed{z = \frac{85-80}{5} = 1 \text{(the score 85 is 1 SD above the mean)}}$$

Key Terms

The **normal distribution** is a symmetric, bell-shaped curve defined by its mean μ and standard deviation σ.

The normal distribution is important both in the real world and in statistics. Many processes and phenomena in the natural sciences and social sciences can be found to have an approximately normal distribution. Importantly for us, we will later see that it is often used in order to make inferences about variables whose true distributions are not known.

Key Results

Empirical Rule: In a normal distribution:

- 68% of the data will be within 1 SD
- 95% of the data will be within 2 SD
- 99.7% of the data will be within 3 SD

Practice: Suppose the scores on a standardized exam are approximately normally distributed with a mean of 500 and a standard deviation of 100.

1. What percentage of students scored between 400 and 600?
2. What percentage of students scored between 300 and 700?
3. What percentage of students scored above 700?

1. The values 400 and 600 are each one standard deviation from the mean:

   $$500 - 100 = 400 \quad \text{and} \quad 500 + 100 = 600$$

   According to the Empirical Rule, approximately 68% of data falls within one standard deviation of the mean.

   $$\boxed{\text{68\% of students scored between 400 and 600.}}$$

2. The values 300 and 700 are each two standard deviations from the mean:

   $$500 - 2(100) = 300 \quad \text{and} \quad 500 + 2(100) = 700$$

   The Empirical Rule says about 95% of the data lies within two standard deviations.

   $$\boxed{\text{95\% of students scored between 300 and 700.}}$$

3. A score above 700 is more than two standard deviations above the mean. Since 95% of scores lie between 300 and 700, the remaining 5% are outside this range. Because the normal distribution is symmetric, half of that 5% is above 700:

   $$\frac{5\%}{2} = 2.5\%$$

   $$\boxed{\text{2.5\% of students scored above 700.}}$$