DATA ANALYSIS ON THE SAT

You should know all of the following terms to have the best chance of performing well on the Data Analysis section of the SAT.

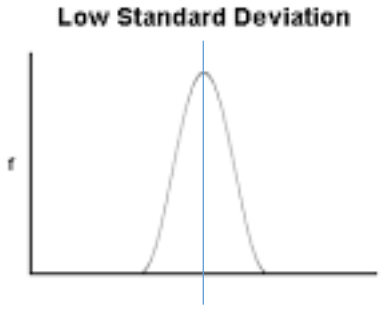
1. **Mean** – the SAT calls it the "arithmetic mean" but another term is the average. To find the average of a data set, you add all of the terms and divide by the total number of terms.

2. **Median** – this is the middle number in a data set when the numbers are arranged from smallest to largest. If there are an even number of terms in the data set, then you average the two middle numbers to get the median.

3. **Mode** – the number that repeats the most in the data set. A data set can have no mode if no numbers repeat. A data set can also have more than one mode if more than one number repeats. *Mode questions are rare on the SAT.*

4. **Standard Deviation** – tells you about the "spread" of the data in a set. Once you know the mean (average) of the data set, the standard deviation tells you how close all of the other terms in your set are to the average.

   For example, in a "normal distribution," most of the terms in the set are close to the average. That distribution looks like this.

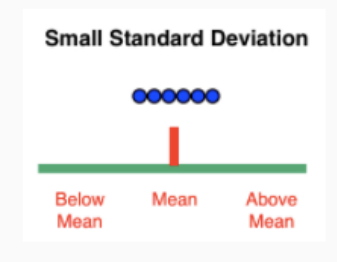


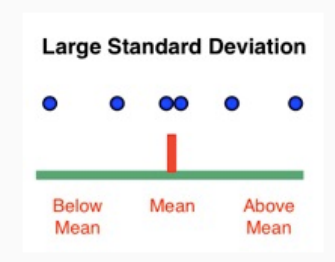
average

In a distribution where a lot of the terms are "clustered" around the mean, your standard deviation is LOW. Data that is closer together, means they deviate from each other less.

The shorter bell-shaped curve suggests that the data points are more spread out and are further from the average. That means the data set will have a higher standard deviation.

Other ways you need to be able to determine standard deviation:





You will not be asked to calculate standard deviation on the SAT. You just have to demonstrate an understanding of what it measures. Remember, you need to know the mean (or be able to approximate it) before you can think about standard deviation.

5. **Confidence Intervals** – when you try to estimate some characteristic about the general population, you have to take a sample and run a study on that smaller amount of the population because we cannot realistically get information on everyone in the world. If the sample is a good representation of the larger population, then our "confidence" in the results is high. If the sample is not a good representation of the general population, then our "confidence" in the results are low.

   For example, if you wanted to know the average height of a high school student, you would just get the heights of the boys in the school. You also wouldn't get the heights of middle school students. Neither of those samples would be a good indicator of the average height of a high school student.

   Also, if you did sample only high school students and included both boys and girls, but you only sampled 10 students out of 1200, the results would probably not be a good representation of the entire school either.

Since we don't know the actual statistical average height of every single student (because we used a smaller sample to get our information), we create a confidence interval. We put the average that we found in the middle and add and subtract some number to the average to create a high value and a low value. We are fairly certain that the "true" average lies somewhere in that interval.

Margin of Error – this can also be called sampling error. This is the amount of error that is expected to occur because we are not taking information from the entire population but instead just asking a smaller sample. The sample results are likely to differ from the actual population results by a certain amount. This is the margin of error and it depends on how precise you want to be.

You will likely be given the SAMPLE mean, the confidence level (percentage) and the MARGIN OF ERROR. Then you will be asked for the CONFIDENCE INTERVAL in which the TRUE MEAN lies.

Example: A random sample of 35 four-door passenger vehicles had a mean gas mileage of 25.9 miles per gallon. The estimate had a margin of error of 2.6 mpg at a 98% confidence level. Of the following, which is the most plausible value for the true mean gas mileage of all four-door passenger vehicles?

A. 24 mpg
B. 29 mpg
C. 32 mpg
D. 35 mpg

To solve this, we take the sample mean (25.9) and add the margin of error (2.6) to it and subtract the margin of error from it to create a confidence interval.

$$25.9 + 2.6 = 28.5$$
$$25.9 - 2.6 = 23.3$$

We are 98% certain that the TRUE MEAN lies between 23.3 and 28.5 mpg. So, the correct choice is A.

Random Sampling: each member of the population has an equal opportunity of being chosen for the sample.

For example, if you wanted to find the average height of everyone in the high school and you only measured the heights of everyone in your homeroom, that might be convenient for you, but it is not a random sample because no one outside of your homeroom had a chance of being sampled.

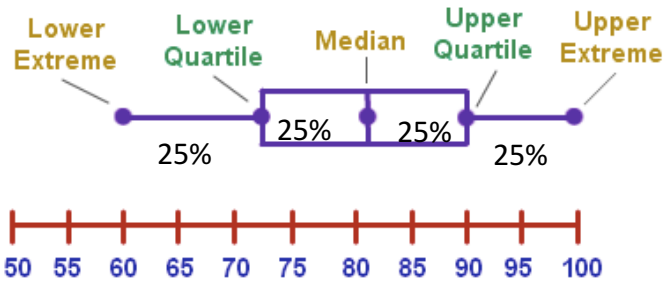
For the results of a study to be able to be generalized to the entire population, random sampling must be used!

Treatment – in a statistical study, the treatment is the variable that is manipulated by the person running the experiment.

For example, if we wanted to know if a new energy drink actually made students get higher grades on an exam, the treatment would be giving the drink to some students (the treatment group) and the other group would not get the drink (control group), and then we would record their exam scores to see who did better.

Remember, correlation does not imply causation!

6. **Box and Whisker Plot**



Shows the median (middle of the data set) NOT THE AVERAGE!

You can estimate the standard deviation from a box and whisker plot.

Each "section" represents 25% of the data in the set.

Histogram – a histogram is like a bar chart, but it puts data into categories (or intervals) and gives you the frequency at which each of the intervals occurs. For example, in the histogram below, the data set contained 30 trees that ranged from 150 to 200 cm in height.