TutorTube: Fundamentals of Statistics

Fall 2020

Introduction

Hello, welcome to another edition of TutorTube, where The Learning Center's Lead Tutors help you understand challenging course concepts with easy to understand videos. My name is Kelly Schmidt, Lead Tutor for statistics at the Learning Center. In today's video, we will explore a few of the fundamental topics and skills that are used in statistics, such as data types, levels of measurement, and summary statistics. Let's get started!

Levels of Measurement

First let's quickly go over some of the basic concepts that are presented in chapter 1. First, **Levels of Measurement**:

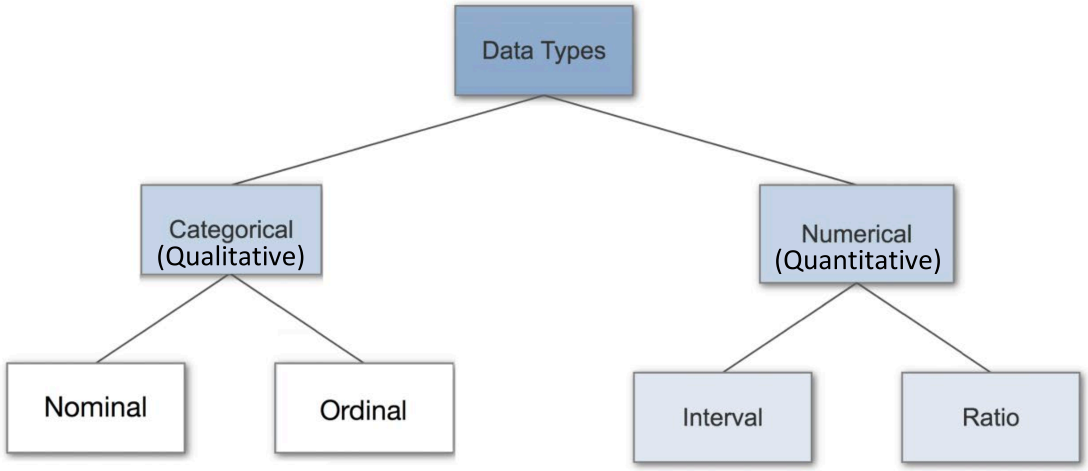


Figure 1: Data Types (Donges, 2018)

All data can be broken down into two basic groups: Categorical (or Qualitative) data and Numerical (or Quantitative) data. Qualitative data represents characteristics, such as eye color, languages spoken, level of satisfaction, and names of people. I think of it as **Qual**itative = **qual**ities.

On the other side, Quantitative data represents numbers or things that can be measured. For this, I think of **Quant**itative = **quant**ities.

Each of these groups can be broken down again until we get to the final four data types that are used in statistics: Nominal, Ordinal, Interval, and Ratio.

Nominal data, much like how it sounds, refer to names. The key difference between Nominal and Ordinal data is that Ordinal data has an inherent order to it. For example, grade levels like Freshman, Sophomore, Junior, Senior follow a logical progression, making them ordinal data. But what about something like languages? If I gave you the data points: English, Spanish, Japanese, and French, would you say they have an inherent order? Since the answer is “no,” we know that these fall into the Nominal category.

On the Quantitative (or Numerical) side, the key difference between Interval and Ratio data types is that Ratio measures have a meaningful zero point. These can be tricky, but the idea is best understood with examples. Ratio measures are what we most commonly think of when we think of measures: height, length, cost, weight. They have a well defined zero point, and we can use math to find their averages, and standard deviations. Interval measure examples are things like time, temperature, SAT scores, or credit scores. They are still numbers, but they don’t necessarily have a well-defined zero point.

Practice:

For practice, try to identify the data types for each of the examples on this page:

Question	***Data Type***
1. What is your gender? o Male o Female o Other	Categorical (Qualitative) - Nominal
2. Percentage grades in a math class	Numerical (Quantitative) - Ratio
3. Social security numbers	Categorical (Qualitative) - Nominal
4. Letter grades in a math class	Categorical (Qualitative) - Ordinal

For the first question, since we are dealing with names instead of numbers, we know that we have some categorical (or qualitative) data. Since there is no order to them, they must be nominal data.

For question 2, we know that percentage grades (like a 90% on a test) are numbers and that it is possible to have a 0 score. Because of this, we classify these as Numerical (Quantitative) Ratio data.

Question 3 can be tricky because social security numbers look like numbers and many people will try to classify them as quantitative data. However, if we think about the context of how SSN's are used, we realize that they actually serve as identifiers, not measures. If I tried to take the average of three strangers' SSN's, the result would tell me nothing. Because of this, we can treat them like names, which makes them nominal data.

For question 4, notice how it differs from question 2. Both measures represent the same concept (grades), but they are measured differently, so we have to treat them differently in statistics. Now that we are only using letter grades, we are limited to A, B, C, D, and F. Since these are not numbers, and they have an order to them, we classify them as ordinal data.

Numerically Summarizing Data – Basic Concepts

Ok, now that we have a good idea of what types of data we'll see in the course, we can move on to another important skill to have: being able to summarize and describe data that you're given.

First, lets talk about some of the symbols and formulas that you'll see throughout the course. In statistics, we have two major groups: populations and samples. Populations are big groups, and we take samples from these big groups in an attempt to understand them better. Measures that relate to the population are usually represented with Greek letters. μ (pronounced myu) is the population mean and σ (pronounced sigma) is the population standard deviation.

$$\text{Population Mean: } \mu = \frac{\Sigma x_i}{N}$$

$$\text{Sample Mean: } \overline{x} = \frac{\Sigma x_i}{n}$$

Figure 2: Formulas for Means (Pearson, 2020)

Population Standard Deviation:

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} = \sqrt{\frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N}}{N}}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n-1}}$$

Figure 3: Formulas for Standard Deviation (Pearson, 2020)

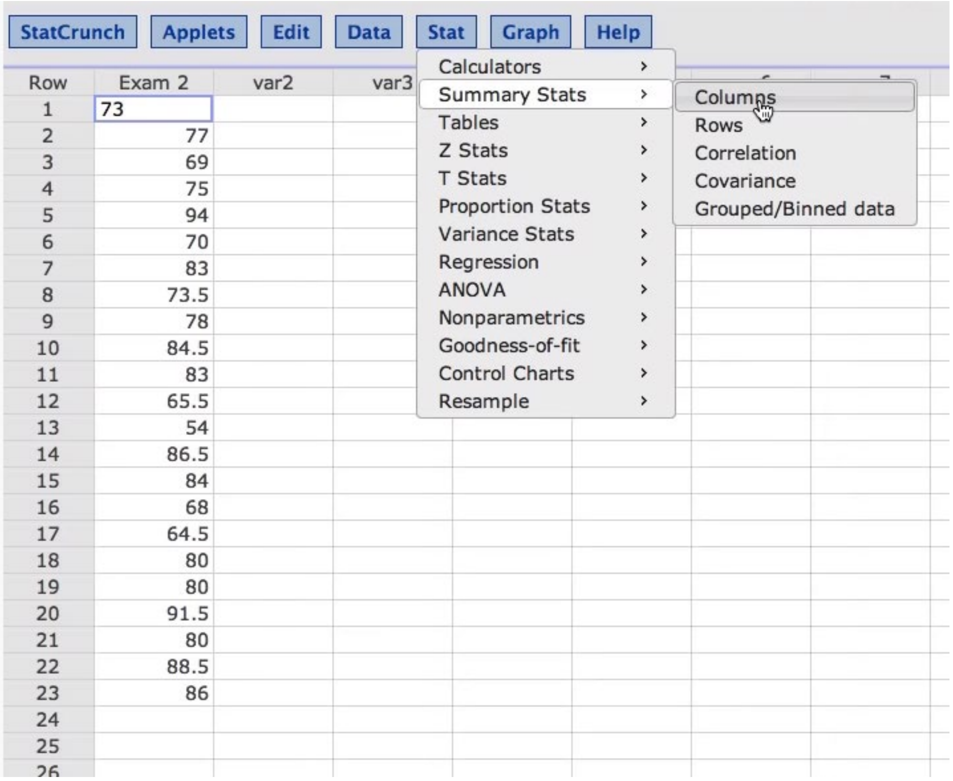
On the other hand, if we are dealing with a sample of a bigger population, we use the symbols $\bar{x}$ for the sample mean and s for the sample standard deviation.

The formulas for calculating each of these may look intimidating, but don't worry about knowing them to the letter. Just know what they represent and when to use them in StatCrunch.

StatCrunch – Summary Statistics

For practice, let's use StatCrunch to find some of these statistics for a random group of numbers, let's say they represent a sample of exam scores (a ratio data type) for a class. I'd encourage you to open your own version of StatCrunch and follow along.

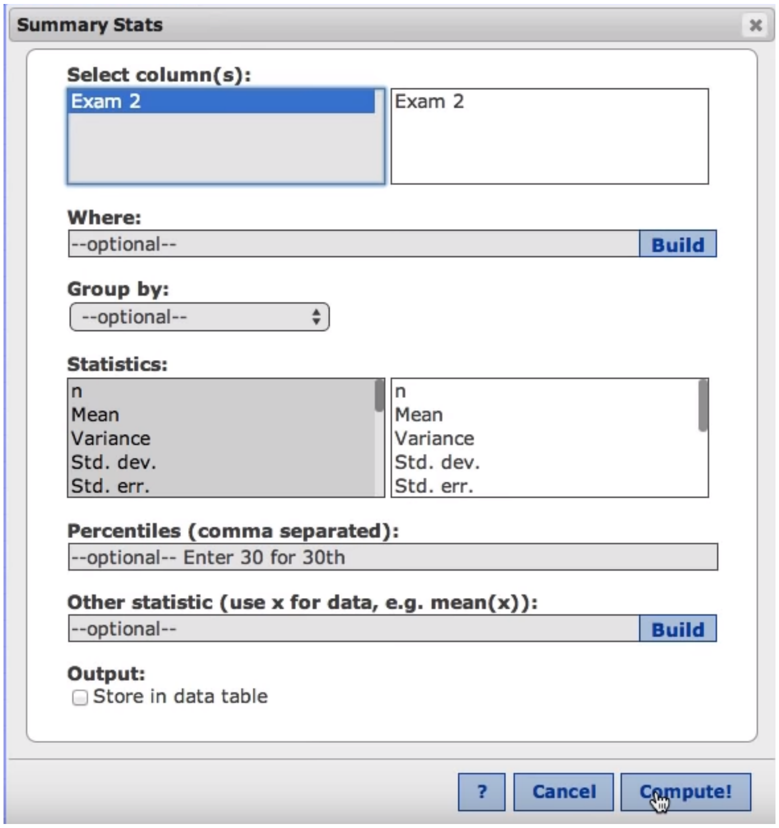
Let's say we have the data entered in a row in StatCrunch. Click on Stat, then Summary Stats, then Columns.



Row	Exam 2	var2	var3
1	73		
2	77		
3	69		
4	75		
5	94		
6	70		
7	83		
8	73.5		
9	78		
10	84.5		
11	83		
12	65.5		
13	54		
14	86.5		
15	84		
16	68		
17	64.5		
18	80		
19	80		
20	91.5		
21	80		
22	88.5		
23	86		
24			
25			
26			

Figure 4: Stat > Summary Stats > Columns

This brings up a new window. Select the name of column that holds the numbers we want to summarize. Here, that's Exam 2. Next, click on Compute.



Summary Stats

Select column(s): Exam 2

Exam 2

Where: --optional--

Group by: --optional--

Statistics:
n
Mean
Variance
Std. dev.
Std. err.

n
Mean
Variance
Std. dev.
Std. err.

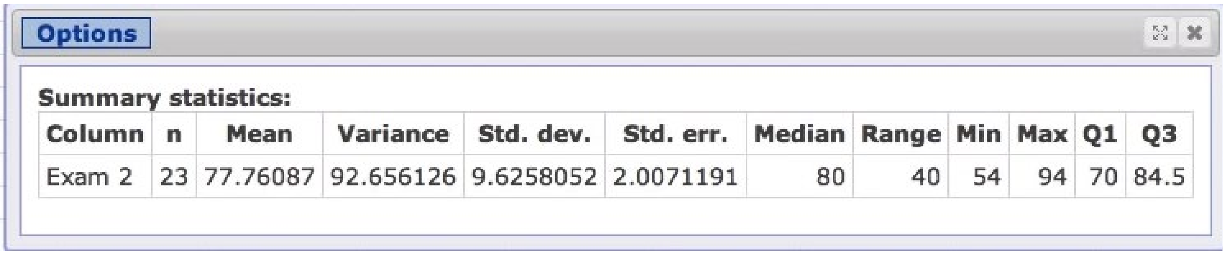
Percentiles (comma separated): --optional-- Enter 30 for 30th

Other statistic (use x for data, e.g. mean(x)): --optional--

Output:
□ Store in data table

Figure 5: Summary Stats Window

Now we see our results window. Notice that we can find the mean of 77.76. Since we said this is a sample, we would represent this data point as $\bar{x}$. The sample standard deviation, listed under Std. dev. Is 9.625. We would represent this statistic with an s.



Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Exam 2	23	77.76087	92.656126	9.6258052	2.0071191	80	40	54	94	70	84.5

Figure 6: Results Window

Notice that we have a number of other descriptive measures here as well: Variance, Median, Range, Min, Max, and Quartiles.

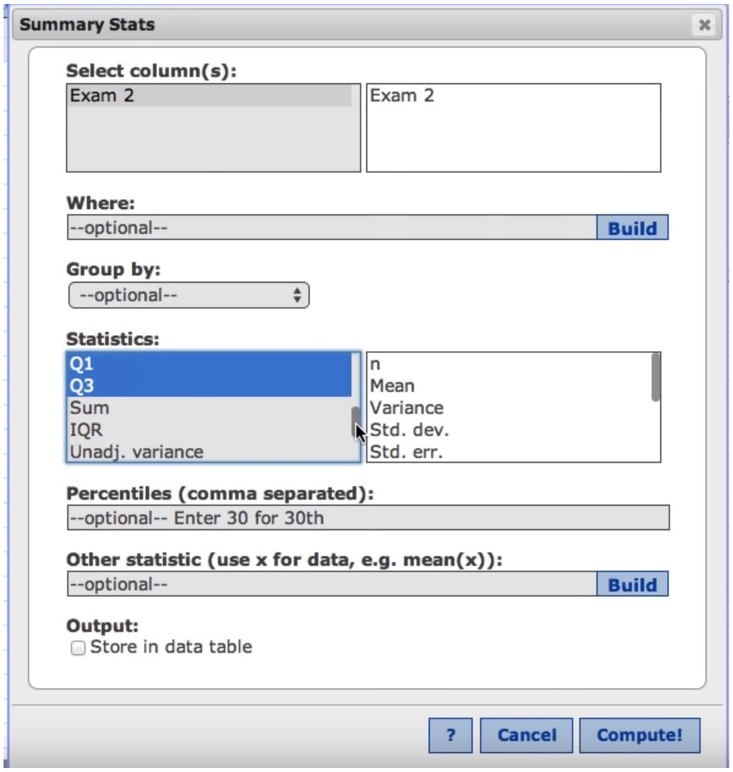
What do these numbers mean? In the context of our example, we had a sample of test scores for some exam. Looking at this results window,

- **n** represents the count of scores, 23. So we had 23 students in this sample.

- **Mean** is the arithmetic average of the scores. The average score on this test was a 77.8%.
- **Variance** is simply the value of the standard deviation squared. If you square 9.626 you will get 92.66. In practice, it tells us how spread out our data is.
- **Std. dev.** is the standard deviation of this sample data. Much like variance, it's also a measure of how spread out the data is. Did most students get scores close to the average of 77%? Or did some get close to zero while others got perfect scores? The standard deviation will tell us this.
- **Std. err**. is the standard error of the data. Don't confuse this with standard deviation! They measure different things. We use standard error when we talk about sampling distributions in chapter 8, so if you are just starting out, don't worry about it.
- **Median** is the middle number in the data set when arranged from smallest to largest. Like mean, it's a way for us to get a sense of the center of the data. (A measure of central tendency)
- **Range** is found by taking the biggest number in the data set and subtracting the smallest number in the data set. So, if the highest score in the class was a 100 and the lowest was a 30, our range would be 70.
- **Min and Max** give us the smallest and largest numbers in the data set. Notice that if you subtract the min (54) from the max (94), you get the range (which is 40).
- **Q1** is quartile 1. If you arranged the scores from smallest to largest, the score at the first quartile would have 25% (or one quarter) of scores below it. Since our first quartile here is at a 70, we can say that 25% of people in the class got worse than a C on this test.
- **Q3** represents the third quartile. Here, if you arranged the scores from smallest to largest, 25% of the scores would be <u>above</u> Q3. A third quartile at 84.5 tells us that only a quarter of the students in this class did better than an 84.5%. Or, phrased another way, ¾ of the class got below an 84.5%.

However, these aren't the only useful statistics we can get from StatCrunch. Let's say that we were curious about the total sum of these scores, or maybe we want to get a better sense of how spread out they are, or which score appeared the most often.

We can click on Options, then Edit to get back to the Summary Stats window.



Summary Stats

Select column(s): Exam 2

Where: --optional--

Group by: --optional--

Statistics: Q1, Q3 (selected); Sum, IQR, Unadj. variance | n, Mean, Variance, Std. dev., Std. err.

Percentiles (comma separated): --optional-- Enter 30 for 30th

Other statistic (use x for data, e.g. mean(x)): --optional--

Output: □ Store in data table

Figure 7: Summary Stats Window - Select Stats

Under "Statistics" we can scroll down to see more options. Use control+click to select the ones that you want your output to display. I am going to select Sum, IQR, and Mode. Now hit Compute again.

This will bring up a new results window with only the statistics we wanted.

- **Mode** is the number that appears most often in the data set.
- **IQR**, the interquartile range, is found by subtracting Q1 from Q3. It is a measure of the range of the middle 50% of our scores.
- **Sum** is the sum of all the data values in our column.

Alright, so with that, we've covered many of the fundamental concepts that will be used throughout the course.

Outro

Thank you for watching TutorTube! I hope you enjoyed this video. Please subscribe to our channel for more exciting videos. Check out the links in the description below for more information about The Learning Center and follow us on social media. See you next time!

References

Pearson. (2020). MyLab: Statistics. *Pearson Higher Education Inc.*

Donges, N. (2018). Data types in statistics. *MachineLearning-Blog.com*. Web.

*All calculations in this video were performed with Pearson StatCrunch 2020 software.