II2202: Research Methodology and Scientific Writing 2011

Introduction to Research and Data Collection Methods

Mark T. Smith
KTH: Swedish Royal Institute of Technology
School of Information and Communication Technology
msmith@kth.se

This talk is about collecting good data

- The ultimate goal is an understanding of how to collect, analyze and use data in an experiment. Specifically:
- The ROLE of data in a scientific study. What it really is for. It is far more important than you may think.
- PLANNING to get the data you need. What kind of data, and how much is enough.
- How to MEASURE data. How to obtain the data you need.
- How to express the SIGNIFICANCE of your data. Showing what your data means.
- You should be able to exercise these skills as an independent scientist, employee, entrepreneur or consultant.

We will be talking about topics in Statistics

- This is not a replacement for a real course in Probability and Statistics.
- However, some basic and important concepts from Probability and Statistics will be covered that are applicable to science in general.
- They will be enough to get you started on meaningful data analysis.
- You will want to supplement them with more concepts that represent standard analysis tools in your discipline area.

Let the Data Speak!

The whole point of data is to *quantitatively* show the value of something.

- ***Show*** how well something works with minimal ambiguity.
- Accurately ***predict*** how well something can or will work.
- Allow people to ***verify*** your work by re-doing it.
- ***Resolve*** selection criteria. Which solution is better based on what?

Other, no less valuable uses:

- Know the preliminary value of a new idea quickly and clearly.
- Establish your credibility. That you know what you are talking about.
- To teach. To pass on knowledge of benefit to others.
- To communicate in a way that others can understand.
- Remove the non-useful elements from technical decisions.

PLANNING, or determining what data you need

Planning is very important. It determines what the design of your study and experiments will be. Wrong experiments, then useless data.

- WHAT is the question you are trying to answer? In other words, what is your *problem statement*?

Examples:

1. “Can a RFID system be designed for a parking solution where the RFID tag can be read within 0.5 seconds when held 4cm or less from the gate reader?” (A technical research statement.)
2. “Would being charged tuition money change a prospective students decision to enroll at the KTH?” (A social research statement.)
3. “Can a personal navigation system be built that can show the position of a person to within 2cm of their actual position?” (Another technical research statement.)

Determining the data you need and the nature of numeric data

Seems easy. Build it, measure it, done! (No, it doesn't work that way.)

Your data has to match your problem statement needs.

- What *resolution* is necessary for your data?
- What *format* should your data be in?
- How do you actually measure, or *sample* the data?
- *How much* data do you need?
- How do you know your data is not *biased*, or a function of something not intended to be measured (noise)?

You answer these as part of your experimental *planning*.

Depending on your field of study, there may be other considerations.

BAD DATA HAS NO SIGNIFICANCE!

Resolution of *Digital* numbers

Resolution relates to the smallest difference in “true” value that a data number can represent. For digital numbers, resolution is reflected in the value of a Least Significant Bit (LSB). For example:

- An 8 bit byte can only resolve something to within one unit out of 256. (That isn’t very much.)
- A 16 bit representation can only resolve something to within one unit out of 65536.
- A 32 bit representation resolves something to within one unit out of 4294967296. A so forth….

- The more LSBs a representation has, then the more resolution you get.

- HOWEVER, this says nothing at all about how actual data is mapped to these representations! You need to be careful!

Number formats and experiments

- Programming languages and data analysis tools like Excel or Matlab have predefined number formats to represent data.
- Use them. You don't need to make up your own, but you can if you want to.
- When using anything predefined, be sure your number formats will accommodate your measurement range and your required resolution.
- Don't state high or excessive resolution if it is meaningless with respect to your data.

Here are some examples.

Example: Range and resolution

Suppose you want to measure temperature. You want:

Measurement range: From -40°C to 120°C.
Resolution: 0.5°C

To represent this, what should you use?

To find out, determine the number of LSBs you need:

$$120-(-40) \Big/ 0.5 = 320 \text{ LSBs}$$

This means you need a data representation having at least:

$$\left(\frac{\log(320)}{\log(2)}\right) = 8.32 \text{ bits}$$

Need more than an 8 bit byte.
A 16 bit representation would be OK.

Example: Excessive resolution

Suppose you read 9 temperature values, and they are:

22, 26, 25, 23.5, 22, 24.5, 27, 26.5, 25

Now suppose you represent these as floating point numbers and you compute the average temperature to obtain:

(22 + 26 + 25 + 23.5 + 22 + 24.5 + 27 + 26.5 + 25) / 9 = 24.611111111111

What does this mean? Not much! Just because your number format can represent a high degree of resolution, that doesn't mean your data does!

Our temperature sensor can only resolve 0.5 degree. So the average of our temperature data readings can imply no higher resolution than that.

Here we can only say that the average temperature is 24.5 degrees.

Getting back to meaningful Data

Now that we know how to represent data with the right resolution, we need to look at how we show what the data really means.

- What is a *measurement*? What is meant by that?
- What is the *accuracy* of your measurements?
- What is the *precision* of your measurements?
- How do these reflect on the quality or the *confidence* one can have in your data?
- How do you make them better, if you need to?

Measurements and other vocabulary

Talking about experiments, data and measurements can be very confusing. We need a standard vocabulary to use.

- An *EXPERIMENT* is a collection of related measurements.
- A *MEASUREMENT* is a data point. Your collection of measurements is what makes up your experimental data.
- Sometimes, people call a measurement a *STATISTIC*.
- A measurement is made by taking *SAMPLES* of something. For example a voltage reading, or someone's opinion. You determine how many samples per measurement you need.
- Note that the *number of samples* you take per measurement is often <u>NOT</u> the same thing as the *number of measurements* you have in an experiment!

Accuracy and Precision

Given a “true” value to measure:

Accuracy relates to the difference between your measurement and the “true” value. It helps to assume you have perfect repeatability when thinking about what accuracy is.

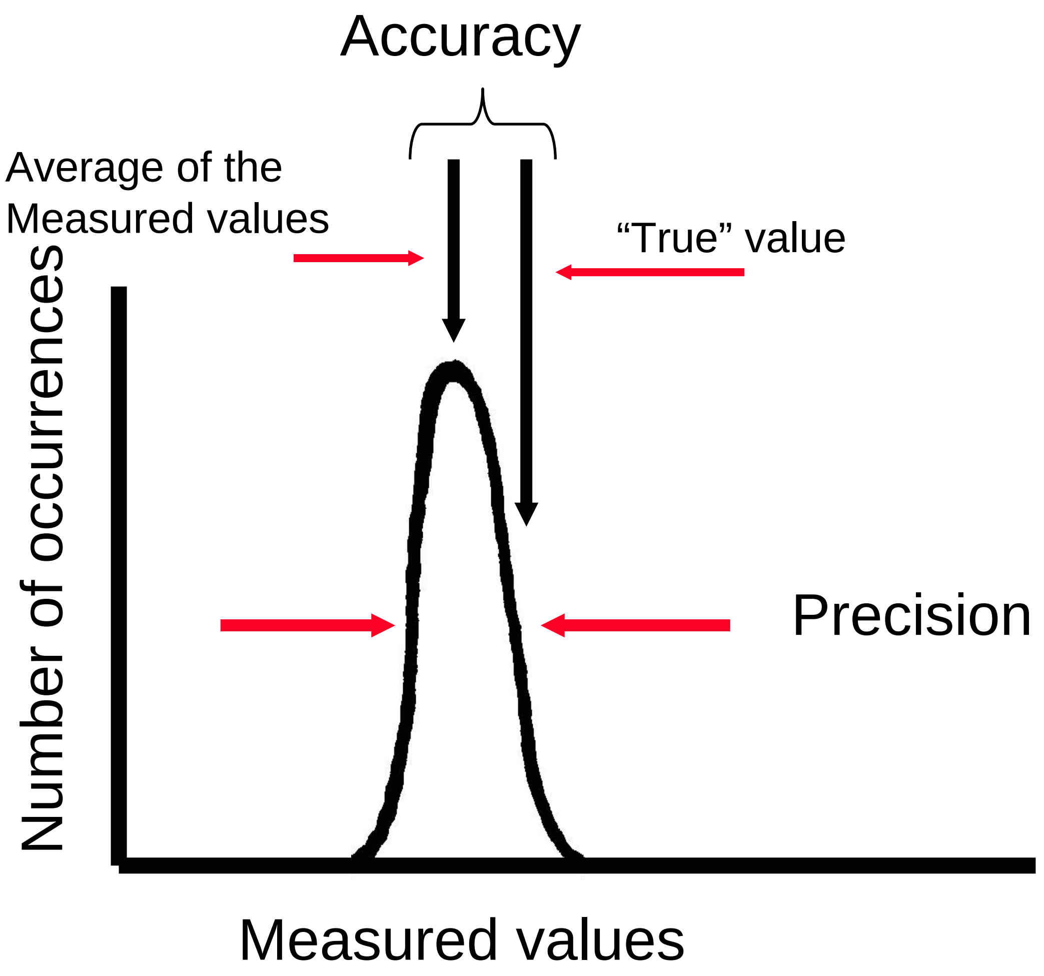
Precision relates to how repeatable your measurements are. It’s possible to be very precise, but not very accurate. It’s also possible for a group of measurements taken together to be quite accurate, but not very precise.

Example: Accuracy and Precision

Accuracy: Is the difference between the "true" value and the average of your actual measurements of the "true" value. Perfect accuracy would result in the average of the actual measurements of the "true" value being exactly the same as the "true" value.

Precision: A measure of the value spread of your actual measurements of the "true" value. Perfect precision would have a spread of zero. The wider the spread, the worse the precision.

Note that it is possible to have good accuracy with bad precision. Also it is possible to have good precision with bad accuracy.



Measurements and meaning

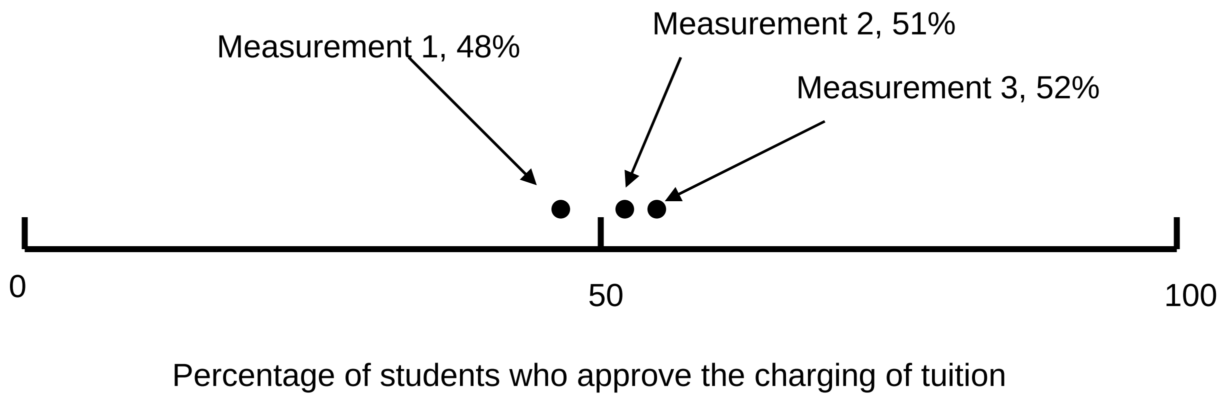
- The reason for making measurements or collecting data is to be able to answer some research question.
- But, how good is your answer? That depends on your data.
- For example, how accurate must data be to be good?
- What about imprecise data? Is that always bad?
- How does one express how good the data is?
- Good data is meaningful data. The key to expressing meaning has its roots in accuracy and precision.

A hypothetical experiment

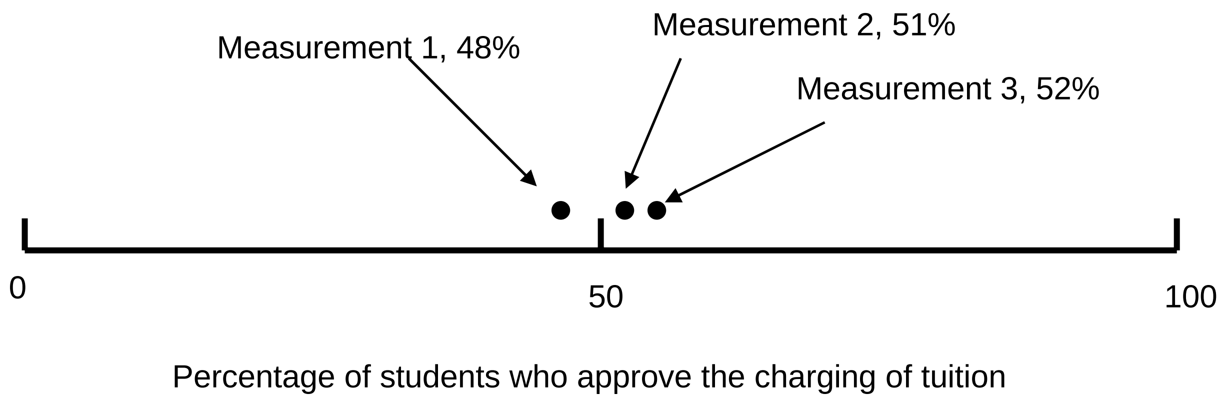
"What percentage of KTH students are in favor of charging tuition for taking KTH courses?"

- There are about 14,500 KTH students total.
- If you could ask them all you would get the "true" answer.
- But, finding and asking them all is not really practical.

- So instead, you go to the list of KTH students and randomly select 100 of them and ask them. You do this 3 times to get 3 measurements.



Estimating the “true” value

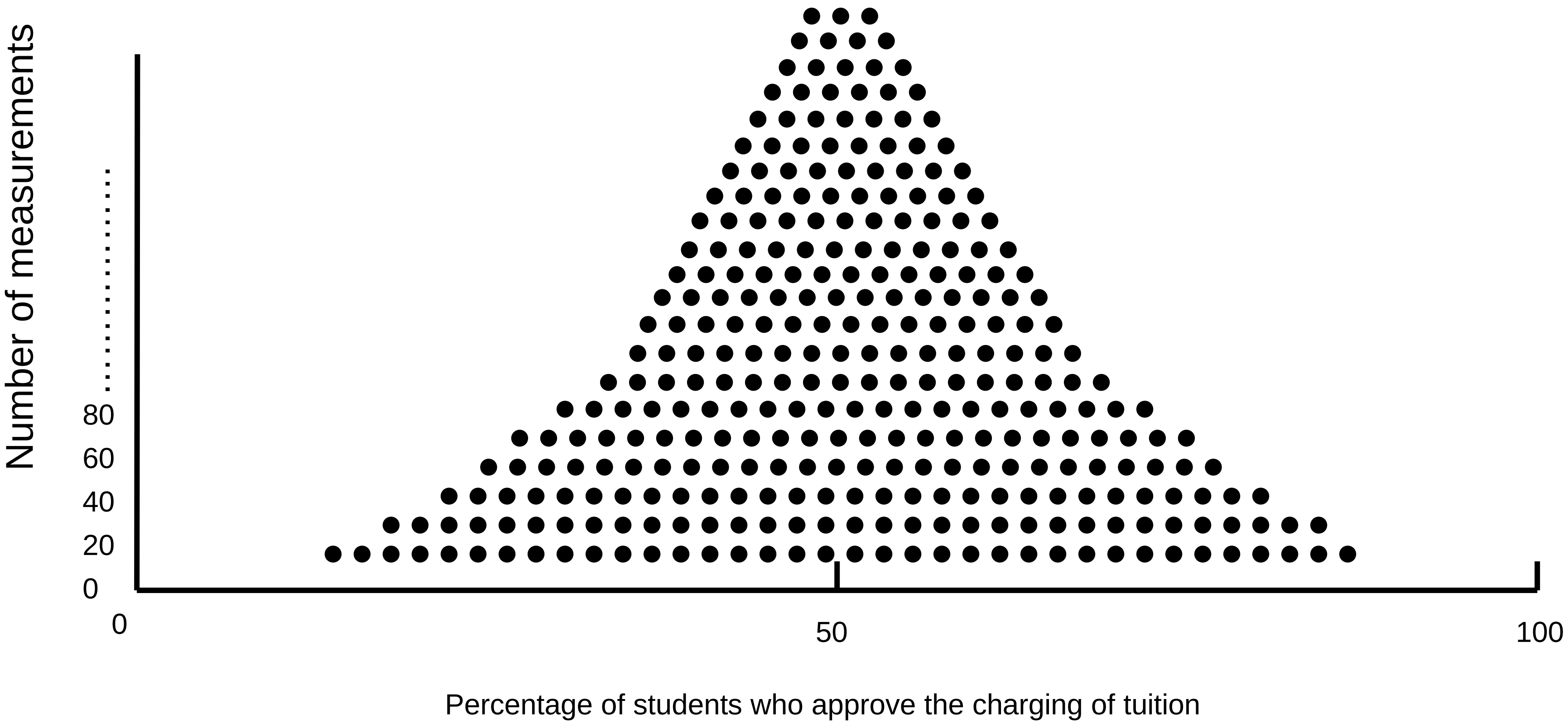


- Each random sample of 100 students give us an estimate of the true value.
- But we have 3 estimates here, and they are all different!
- They can’t all be right, so clearly we do not have perfect accuracy.
- They are all different, so we don’t have perfect precision either.

- How can we use this to estimate what the “true” value really is? To see that, let’s get a lot more 100 student samples and plot them.

A lot more samples

As we take more measurements, the data starts to show meaningful things. The values with the highest number of measurements is probably nearer to the “true” value. The shape of the data is also interesting.



Mean and standard deviation

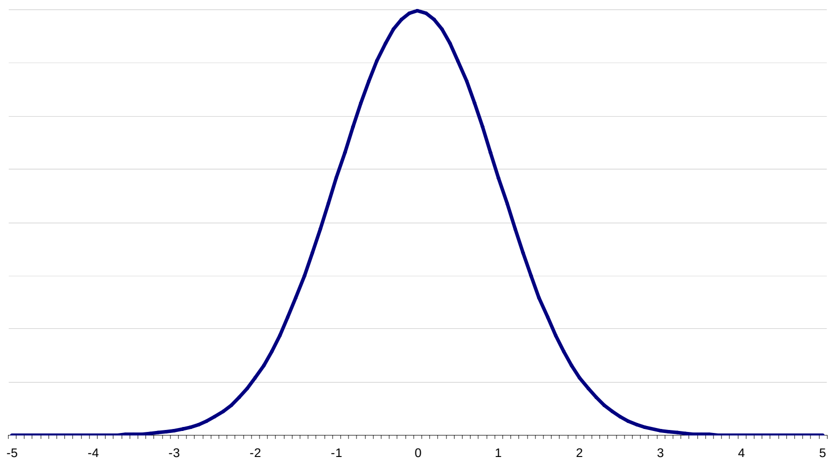
The mean is the average of your measurements. It is an estimation of the "true" value you are trying to find out. It reflects the accuracy of the data.

The standard deviation is the average of how the measurements differ from the mean. It reflects the precision of the data.

$$\mu = \frac{1}{N}\sum_{i=0}^{N-1} k_i$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=0}^{N-1}(k_i - \mu)^2}$$

$\text{Mean} = 0,\ \sigma = 1$



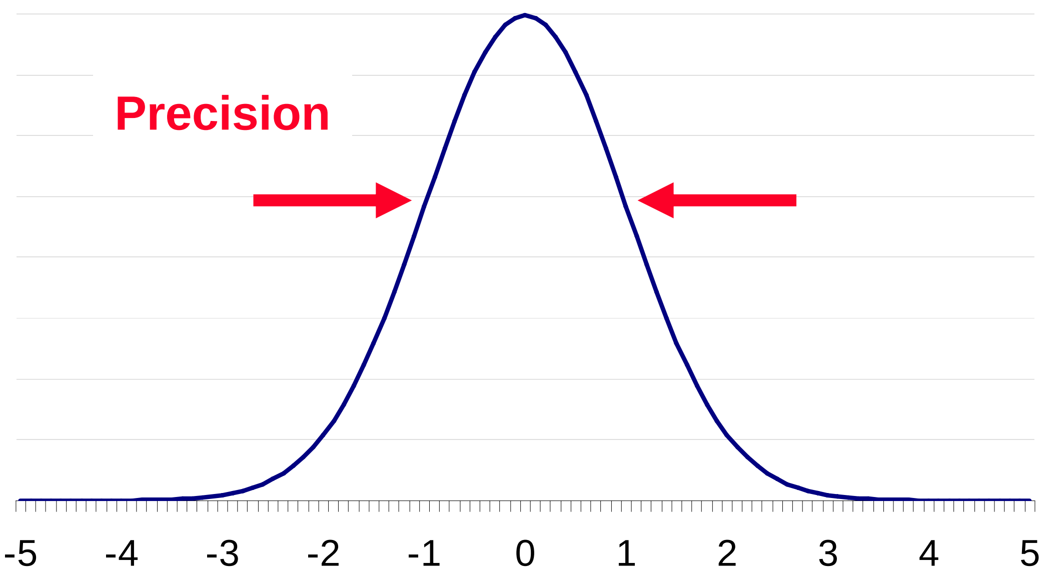
N = total number of *measurements*
k = *measurement* value

Note that N is <u>not</u> the number of samples per measurement!

Precision of the measured data

You can see now the relationship between standard deviation and precision. The larger the standard deviation, the worse the precision.

$$\text{Mean} = 0,\ \ \sigma = 1$$



The reason this is important is that it allows you to predict how much confidence you have that your data is within a certain distance of the “true” value. In other words, it is related to how good your data is.

Standard Deviation and measurement error

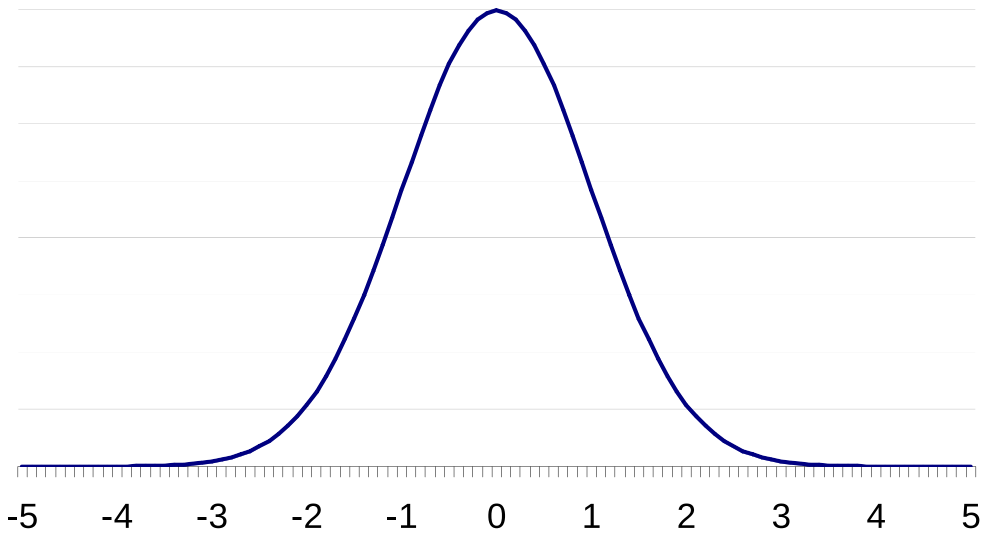
- The standard deviation is the average of how the measurements differ from the mean. That amount is ONE standard deviation.
- Lets say in our experiment about the question of tuition that we get the result that the mean=0.5 and the SD = 0.025.
- In other words, the experiment estimates that 50% of the students agree that the KTH should charge tuition. One SD is 5% of this. Our estimate is not precise, and the amount that it is off averages to 5%.
- This 5% that is our one SD is also called our *standard error*.

Standard error

- Standard error gives you a measure of how good your data is because it tells you how reliably a single measurement will estimate the “true” value.
- In our experiment, it says that one measurement consisting of 100 randomly sampled students will on average be wrong by 5% of the experiment mean value.
- This is OK, but it really isn’t very useful. We don’t want to know how wrong the experiment will be.
- We want to know how confident we can be that the experimental data reflects the “true” value, and by how much.

Properties of normal distributions

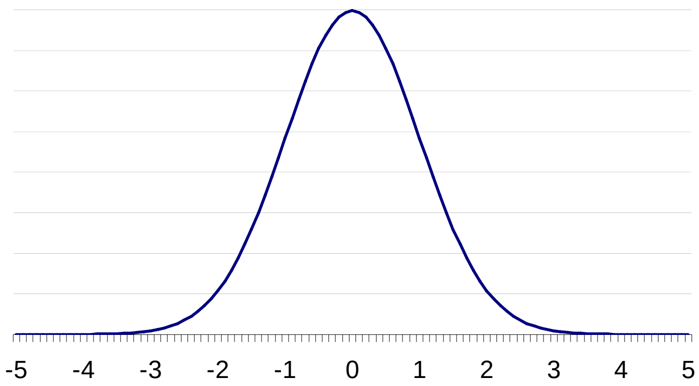
$$\text{Mean} = 0, \ \sigma = 1$$



- The probability that a measurement value for this experiment will exist anywhere under the normal curve is 1.
- The area accounted to each SD is also consistent in a normal curve. This area is equal to the probability that a measurement will fall within one SD away from the mean or “true value”.
- To determine that, we need to know the area of the curve between the mean and one SD.

You can compute areas, or look them up

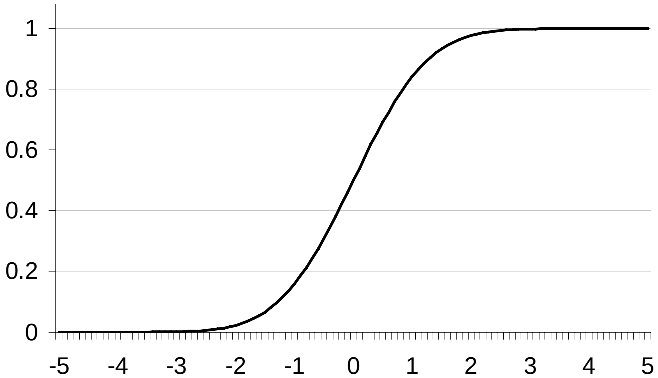
$$\text{Mean} = 0,\ \sigma = 1$$



$$F(x) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} \exp\left[\frac{-(u-\overline{X})^2}{2\sigma^2}\right] du$$

Normalizing $\overline{\mathrm{X}} = 0 \text{ and } \sigma = 1$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-u^2/2} du$$



This is called the Cumulative Distribution Function (CDF) of the normal curve. From this you can take off numbers that give you the probability of being in some range of the normal curve.

Using the CDF to determine probability

The probability of a measurement value being within 1 SD of the mean:

$$P = \Phi(1) - \Phi(0)$$

$$= 0.8413 - 0.5$$

$$= .3413 \text{ or } 34\%$$

The probability of a measurement value being within +-1 SD of the mean:

$$P = \Phi(1) - \Phi(-1)$$

$$= 0.8413 - 0.1587$$

$$= .6826 \text{ or } 68\%$$

x	$\Phi(x)$	x	$\Phi(x)$
-3.4	0.0003	0	0.5000
-3.3	0.0005	0.1	0.5398
-3.2	0.0007	0.2	0.5793
-3.1	0.0010	0.3	0.6179
-3	0.0013	0.4	0.6554
-2.9	0.0019	0.5	0.6915
-2.8	0.0026	0.6	0.7257
-2.7	0.0035	0.7	0.7580
-2.6	0.0047	0.8	0.7881
-2.5	0.0062	0.9	0.8159
-2.4	0.0082	1	0.8413
-2.3	0.0107	1.1	0.8643
-2.2	0.0139	1.2	0.8849
-2.1	0.0179	1.3	0.9032
-2	0.0228	1.4	0.9192
-1.9	0.0287	1.5	0.9332
-1.8	0.0359	1.6	0.9452
-1.7	0.0446	1.7	0.9554
-1.6	0.0548	1.8	0.9641
-1.5	0.0668	1.9	0.9713
-1.4	0.0808	2	0.9772
-1.3	0.0968	2.1	0.9821
-1.2	0.1151	2.2	0.9861
-1.1	0.1357	2.3	0.9893
-1	0.1587	2.4	0.9918
-0.9	0.1841	2.5	0.9938
-0.8	0.2119	2.6	0.9953
-0.7	0.2420	2.7	0.9965
-0.6	0.2743	2.8	0.9974
-0.5	0.3085	2.9	0.9981
-0.4	0.3446	3	0.9987
-0.3	0.3821	3.1	0.9990
-0.2	0.4207	3.2	0.9993
-0.1	0.4602	3.3	0.9995
0	0.5000	3.4	0.9997

A few more useful Gaussian properties

- The probability of a measurement value being within +-1 SD of the true value is 68%.
- The probability of a measurement value being within +- 2 SD of the true value is 95%.
- The probability of a measurement values being within +- 3 SD of the true value is 99.9%
- This is true of ALL Gaussian distributions!

Confidence

- Now, we can say something useful about our data. We can say:

"I am 95% confident that that between 40% and 60% (+- 2 SD) of students approve of the idea of charging tuition."

- This is useful and valuable because we have specified two new components:

- *Confidence interval*: A range of values within which a measurement value is estimated to be. In our case our confidence interval is within 2 SD (plus or minus) or +- (2 X 5%) of the true value.

- *Confidence level*: The estimated probability that a measurement value is within the stated confidence interval. In our case our confidence level is .95 or 95%.

Increasing the precision

- In our example, I claimed:

"I am 95% confident that that between 40% and 60% (+- 2 SD) of students approve of the idea of charging tuition."

- Sounds nice, but 40% to 60% is not all that precise. How can I get better precision?

- Recall that 1 SD in our study is 5%. It was determined by solving the equation for standard deviation.

- Let's look again at the formula and see what makes it work.

Better precision

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=0}^{N-1}(k_i - \mu)^2}$$

- u is the estimation of the true value.
- N is the number of measurements. It looks like increasing N may help, but not really. The key is the relationship between k and u.
- k is a measurement. The closer its value is to u then the smaller the SD.
- Remember that k is itself a mean. It is the mean of the 100 samples used in each measurement!
- So, by increasing the sample size, k will approach the value of u!

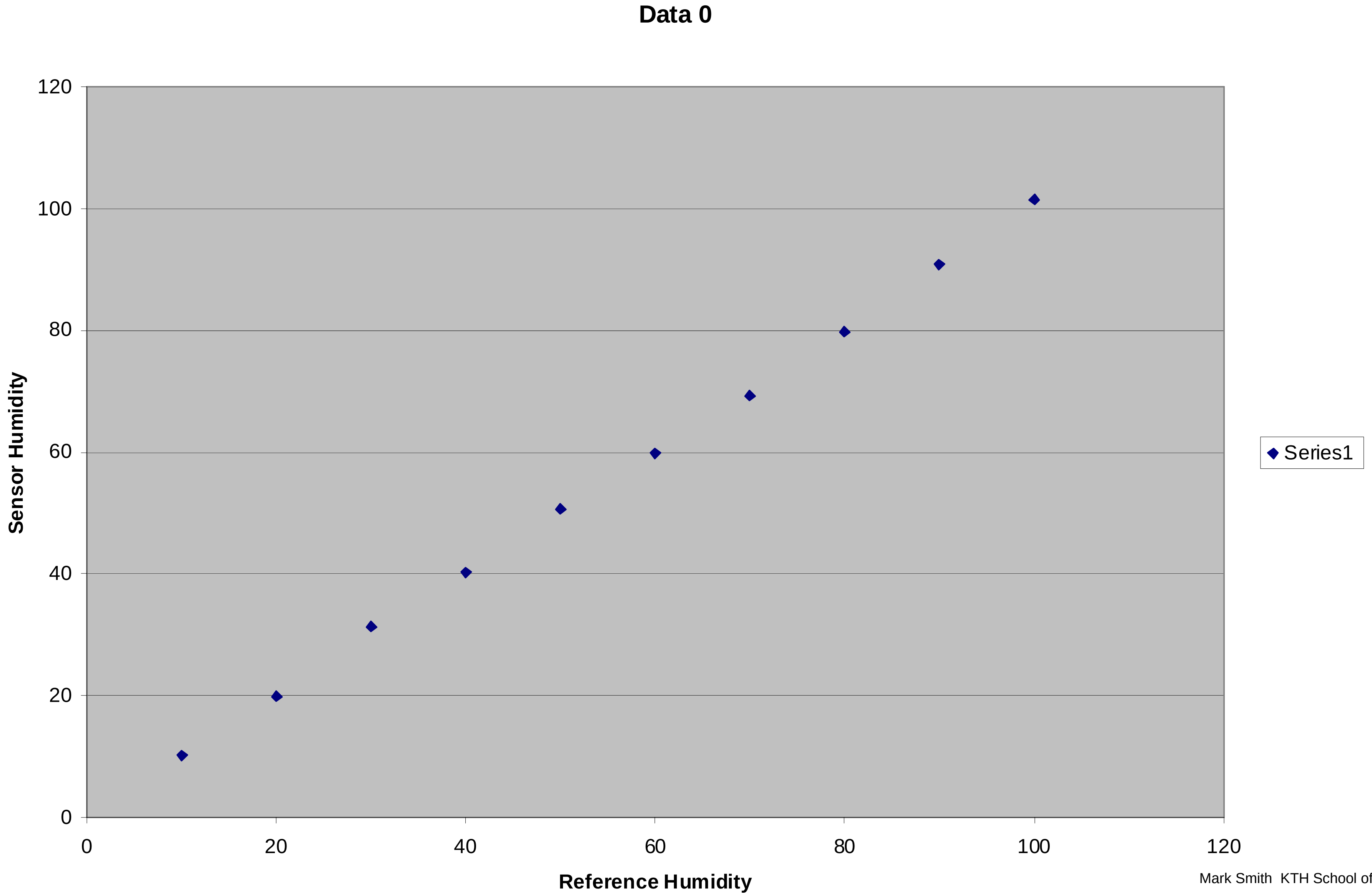
Better precision

$$k = \frac{1}{S}\sum_{i=0}^{S-1} p_i \text{ and } \mu = \frac{1}{N}\sum_{i=0}^{N-1} k_i$$

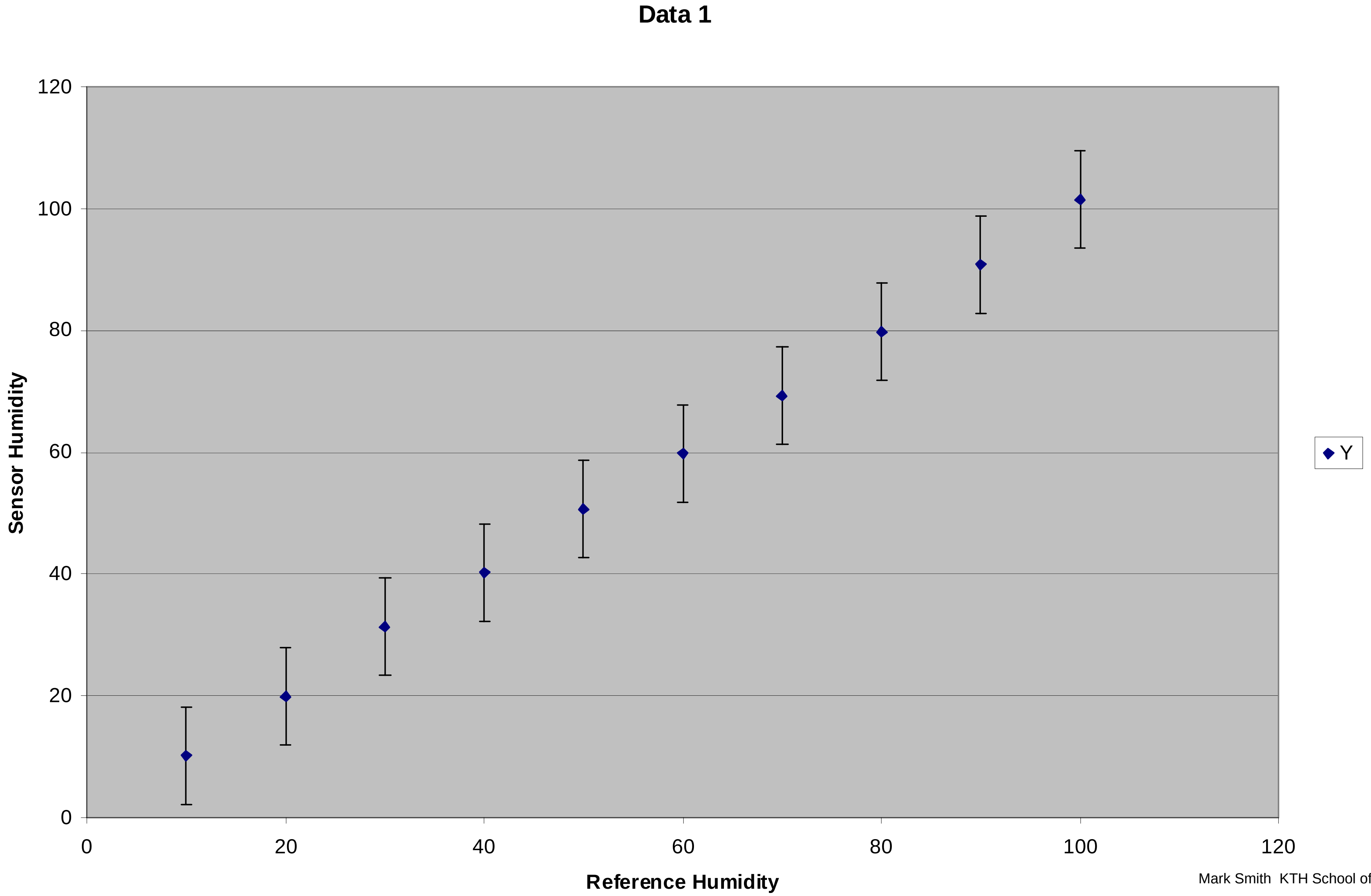
- S is the sample size, and N is the number of measurements.
- p is the value given by each person who is sampled in the study.
- As S gets larger and larger, k starts to approach u, the true mean.
- That means that as S gets bigger, the difference between u and each measurement (k_i) gets less.
- Standard deviation goes down, and precision goes up.
- Choose your sample size for the precision/error tolerance you need.

Example use: A Humidity Meter

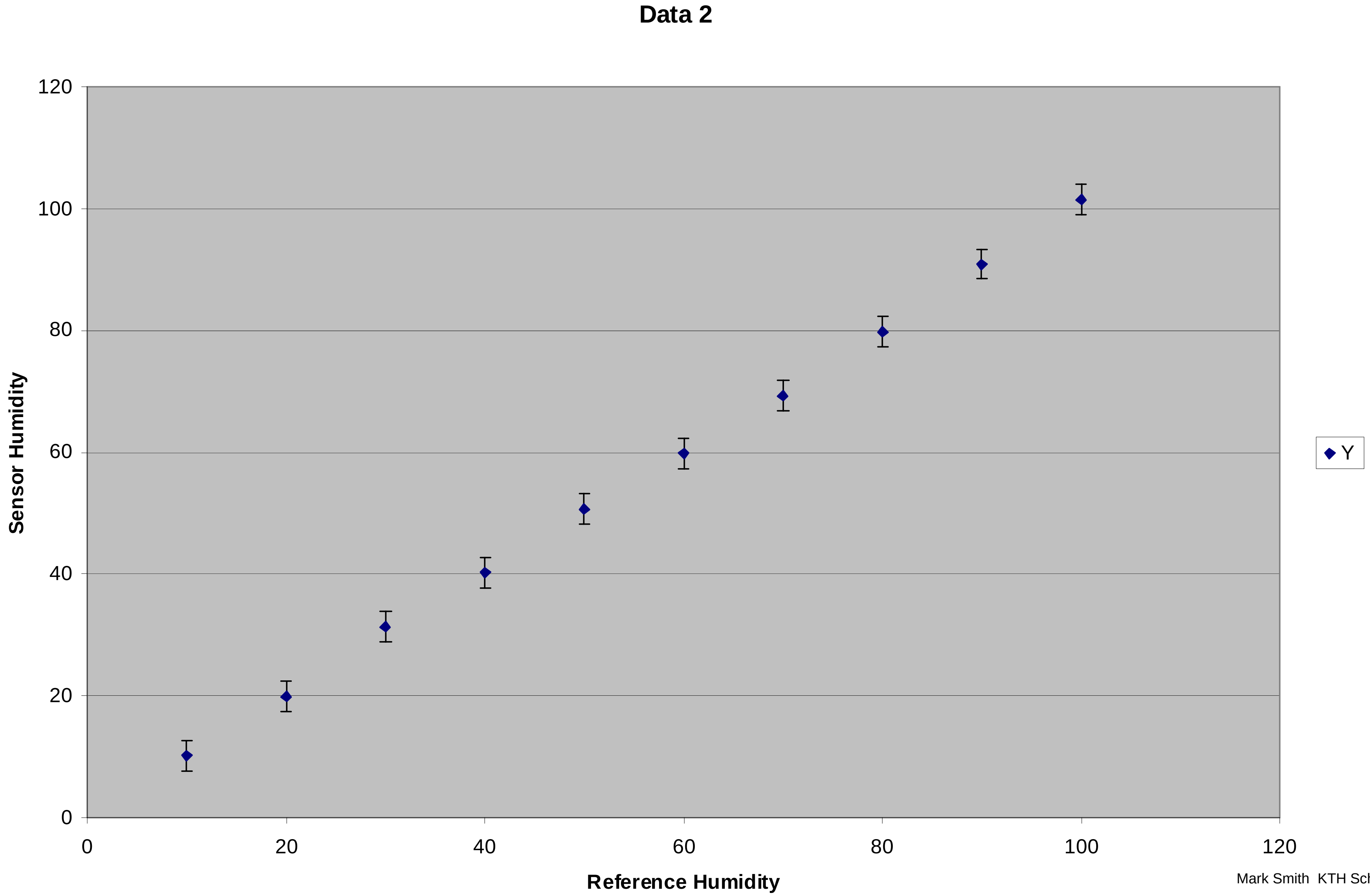
Reads humidity from 10% to 100% with a resolution of 10%



This says a lot more
(and shows a problem too!)

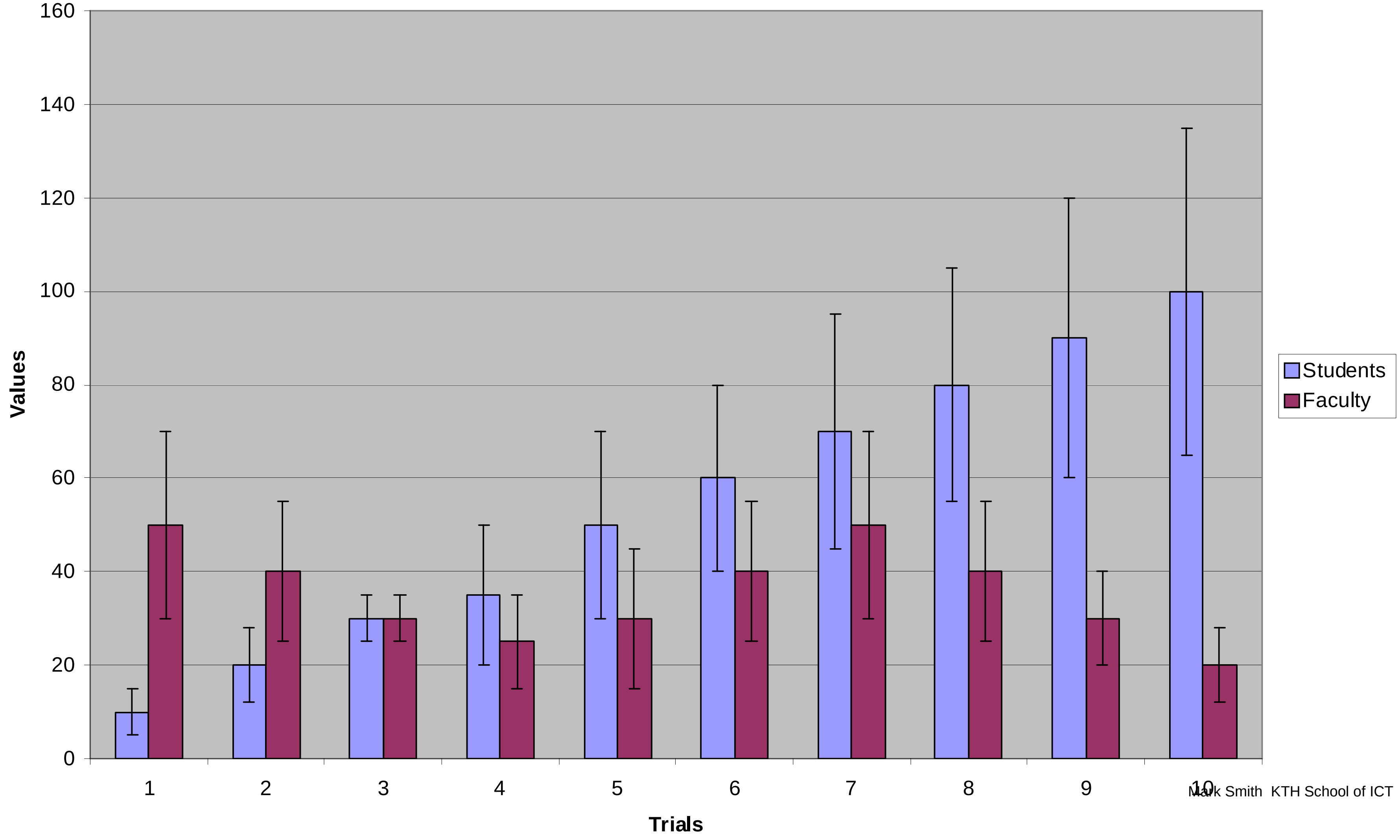


With lower Standard Error (larger sample size per measurement)



Example: Histograms

Where can you see evidence of confidence of a significant difference?



More tools

- We have seen how resolution, measurements and precision can work together to allow you to interpret the meaning of your data.
- This is done by using Gaussian distribution properties to help analyze how your data varies with respect to the “true value”.
- They also allow you to express where you are confident about the meaning of your data (confidence interval) and the degree to which you are confident (confidence level). This establishes the significance of your data.
- There are many other statistical tools that allow you to analyze how your data varies, and establish its significance. For example one and two way Analysis of Variance (ANOVA).
- Find them. Use them.