Statistical Methods For Engineers

ChE 477 (UO Lab)

Larry Baxter & Stan Harding

Brigham Young University

Deductive vs. Inductive Reasoning

- **Deductive Reasoning**
  - **Draw specific conclusions based on general observations.**
  - **Second nature to most physical science and engineering communities.**
  - **Commonly grounded in general physical laws and lends itself to logical analyses/diagrams.**
- **Inductive Reasoning**
  - **Draw general conclusions based on specific observations.**
  - **Frequently abused by both technical and lay communities (component of bigotry, prejudice, and narrow mindedness).**
  - **Statistics provides quantitative and defensible basis for such analysis.**

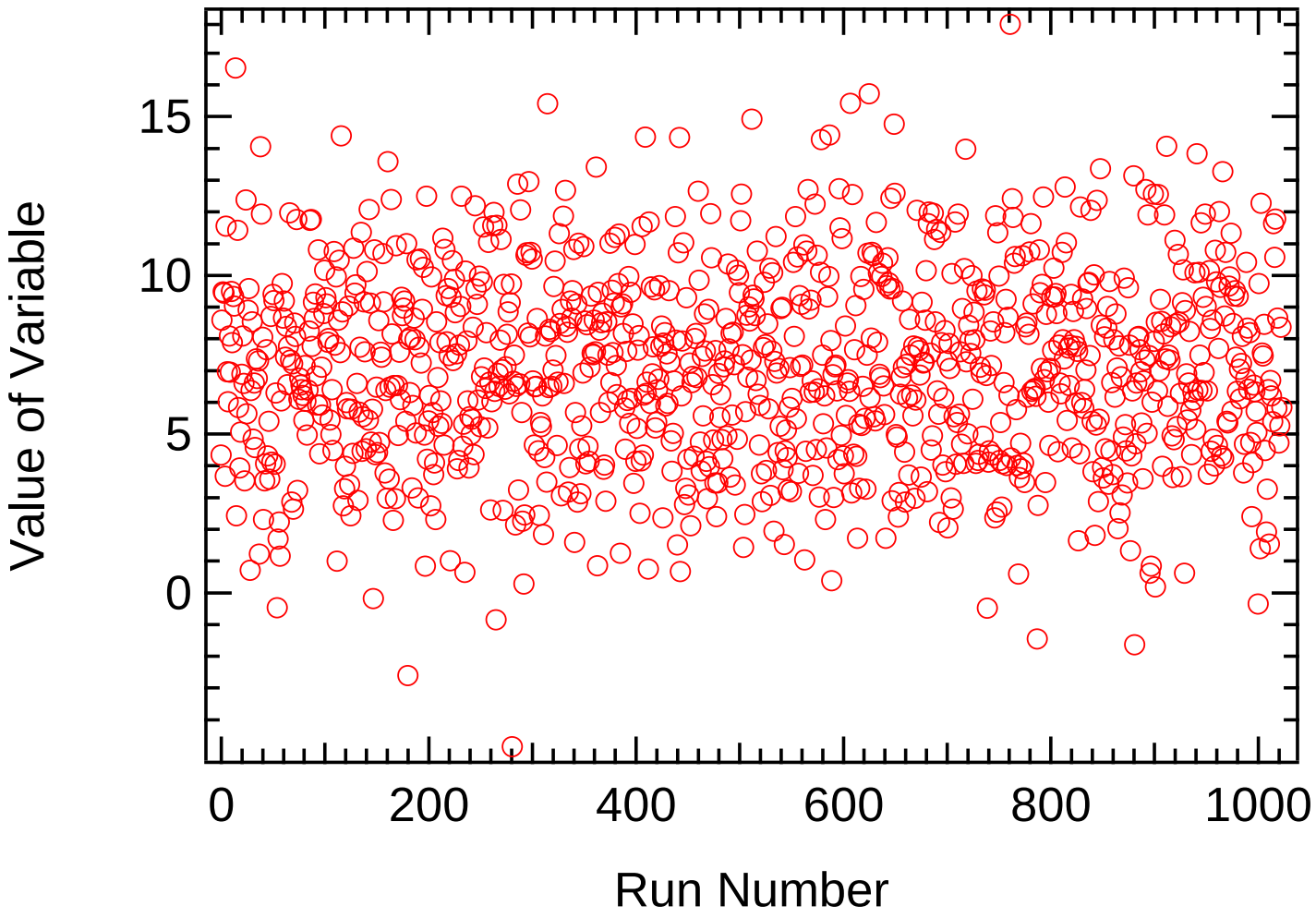
Population vs. Sample Statistics

- **Population statistics**
  - **Characterizes the entire population, which is generally the unknown information we seek**
  - **Mean generally designated μ**
  - **Variance & standard deviation generally designated as σ^2, and σ, respectively**
- **Sample statistics**
  - **Characterizes a random, hopefully representative, sample – typically data from which we infer population statistics**
  - **Mean generally designated $\bar{x}$**
  - **Variance & standard deviation generally designated as s^2 and s, respectively**

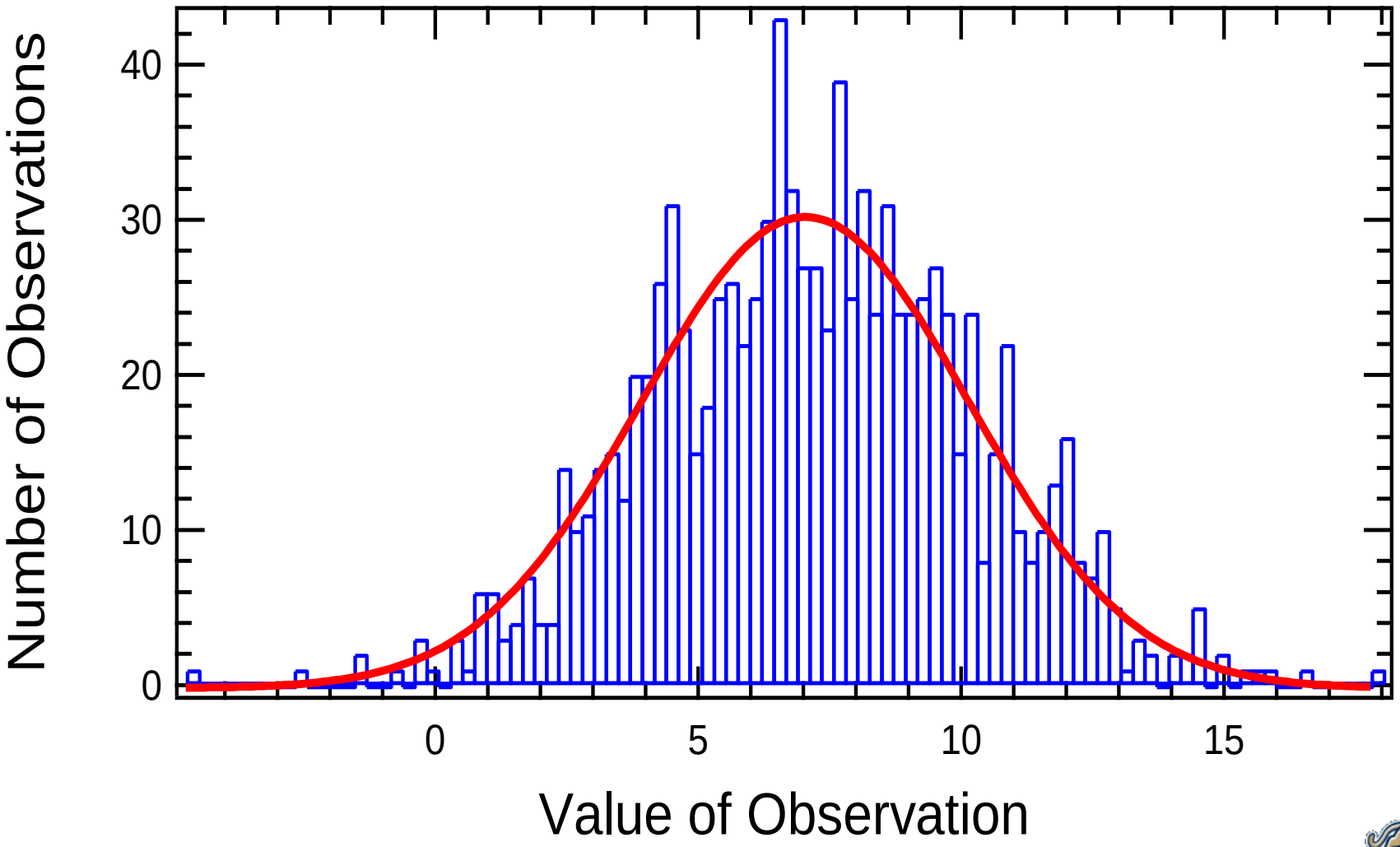
Overall Approach

- **Use sample statistics to estimate population statistics**
- **Use statistical theory to indicate the accuracy with which the population statistics have been estimated**
- **Use trends indicated by theory to optimize experimental design**

Data Come From pdf

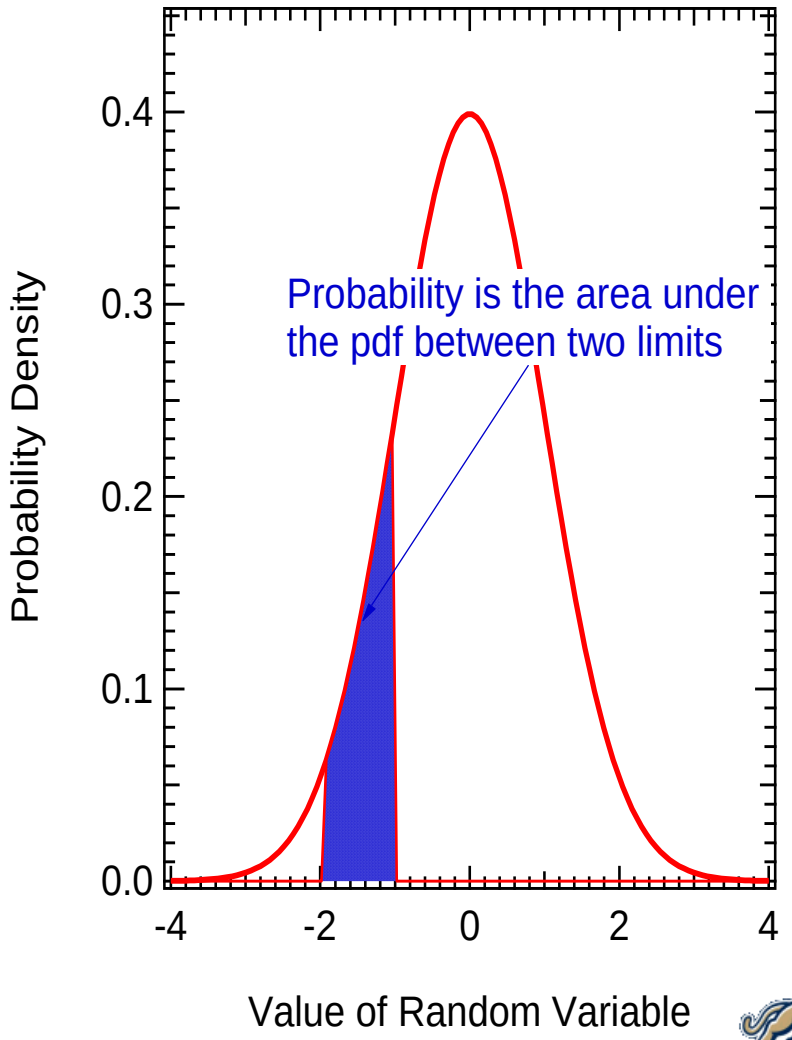


Histogram Approximates a pdf

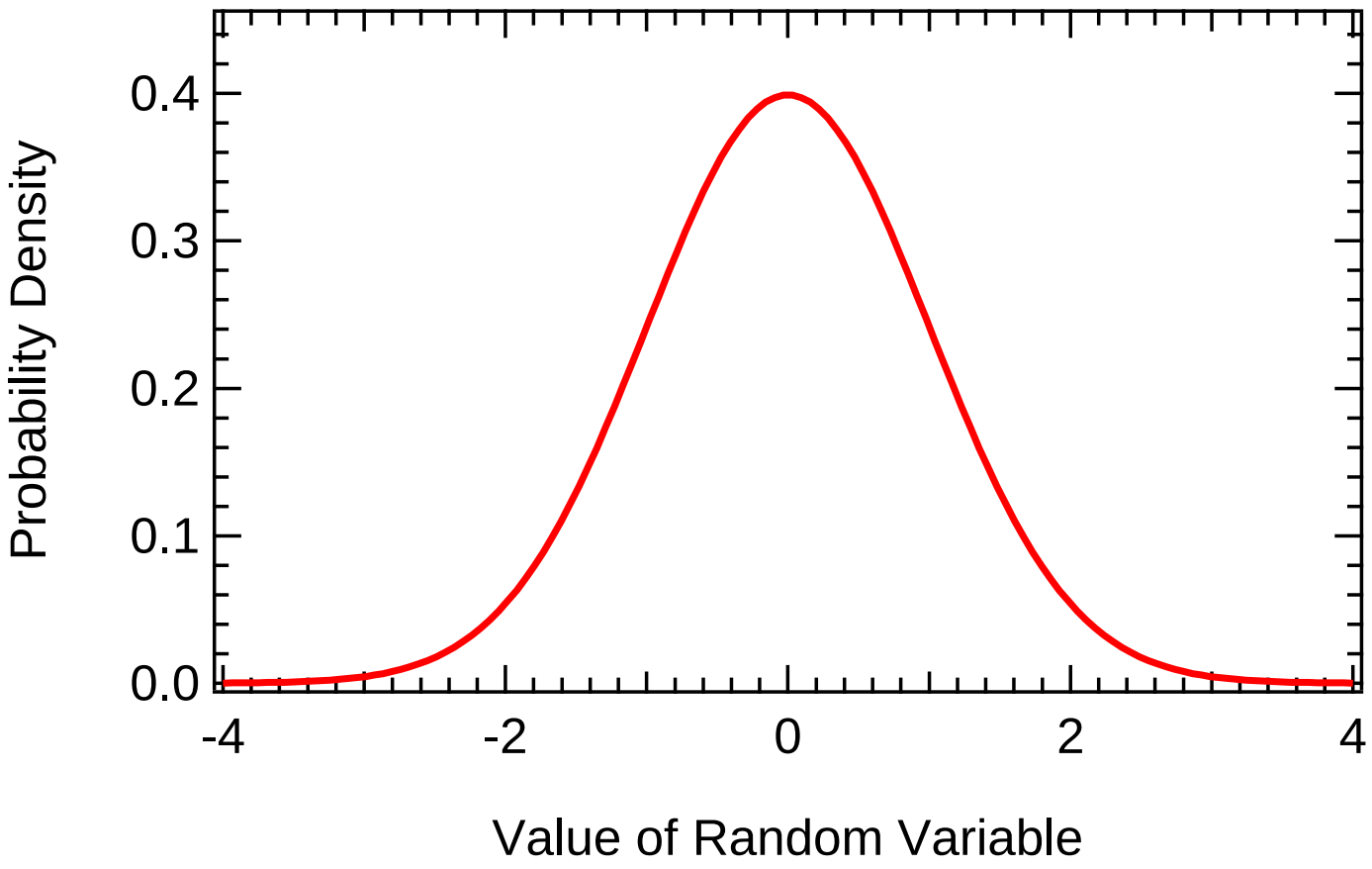


All Statistical Info Is in pdf

- **Probabilities are determined by integration.**
- **Moments (means, variances, etc.) Are obtained by simple means.**
- **Most likely outcomes are determined from values.**



Gaussian or Normal pdf Pervasive

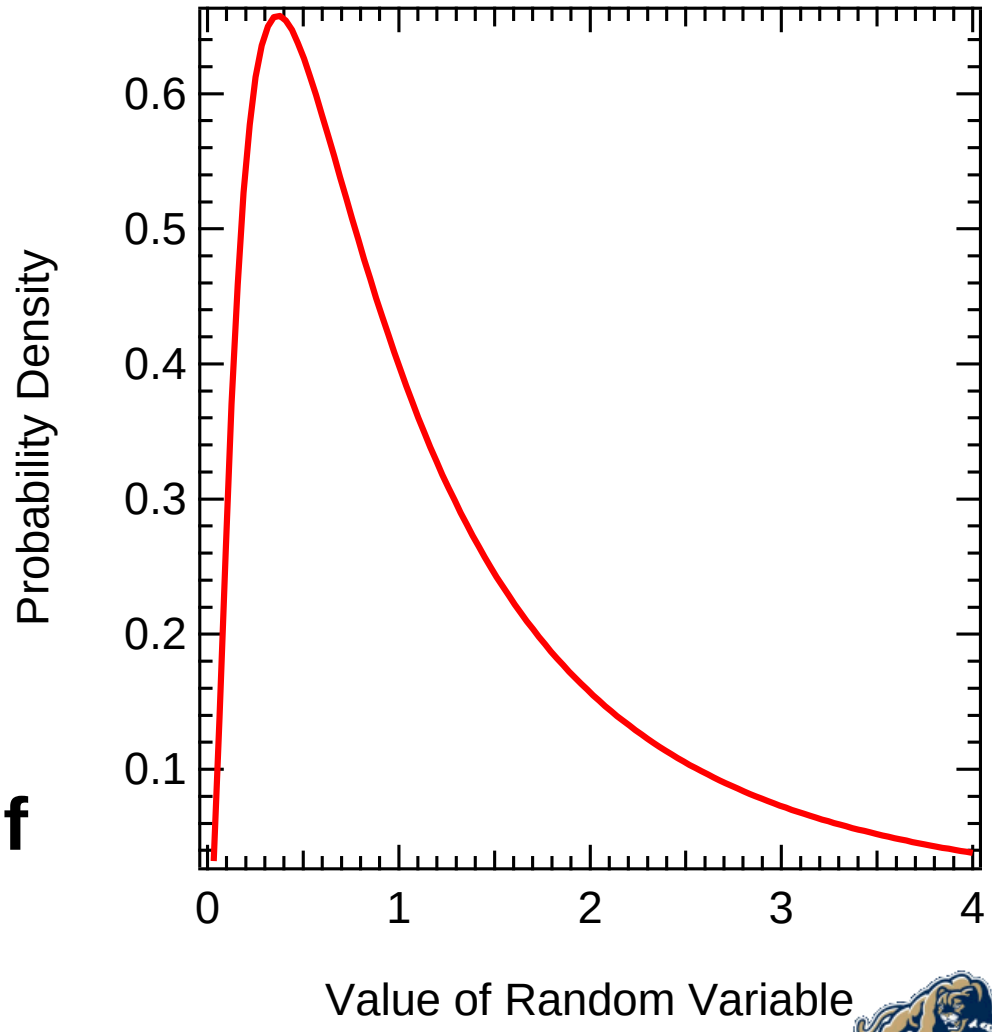


Properties of a Normal pdf

- About 68.26%, 95.44%, and 99.74% of data lie within 1, 2, and 3 standard deviations of the mean, respectively.
- When mean is zero and standard deviation is 1, it is referred to as a standard normal distribution.
- Plays fundamental role in statistical analysis because of the Central Limit Theorem.

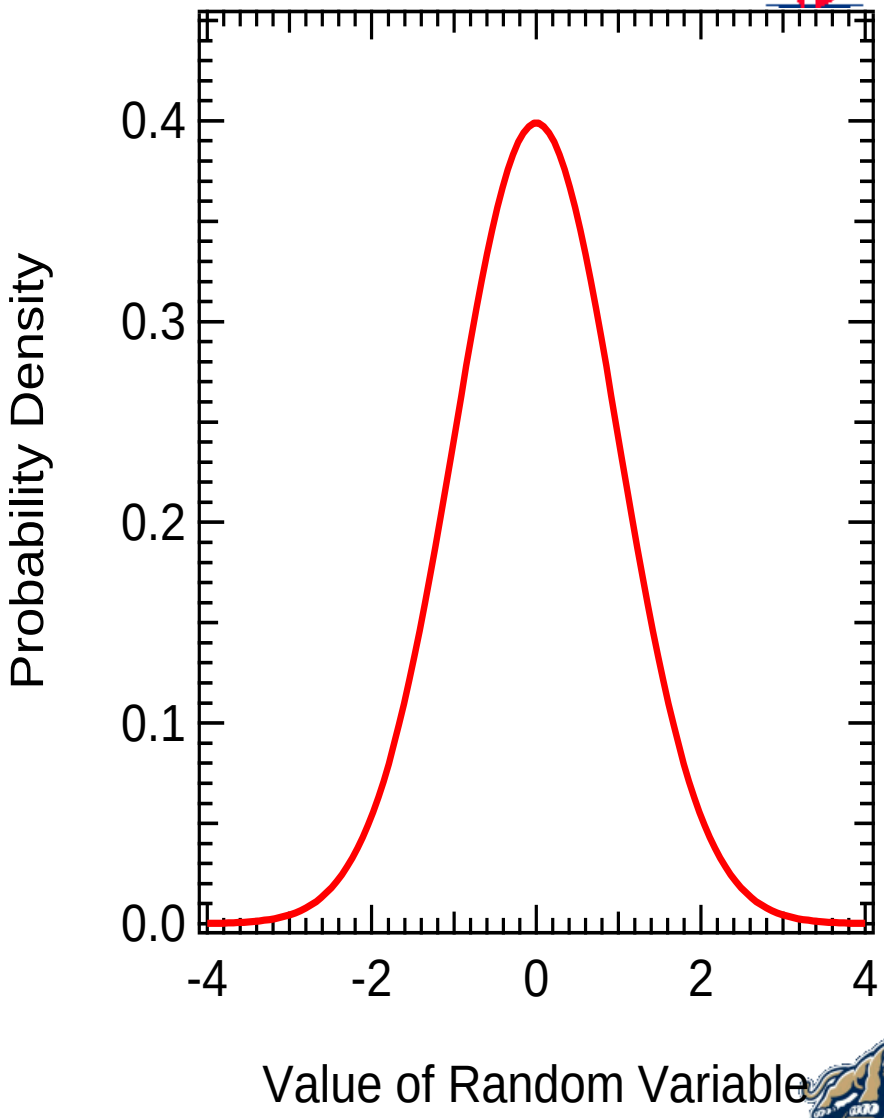
Lognormal Distributions

- **Used for non-negative random variables.**
  - **Particle size distributions.**
  - **Drug dosages.**
  - **Concentrations and mole fractions.**
  - **Duration of time periods.**
- **Similar to normal pdf when variance is < 0.04.**



Student's t Distribution

- **Widely used in hypothesis testing and confidence intervals**
- **Equivalent to normal distribution for large sample size**
- ***Student* is a pseudonym, not an adjective – actual name was W. S. Gosset who published in early 1900s.**

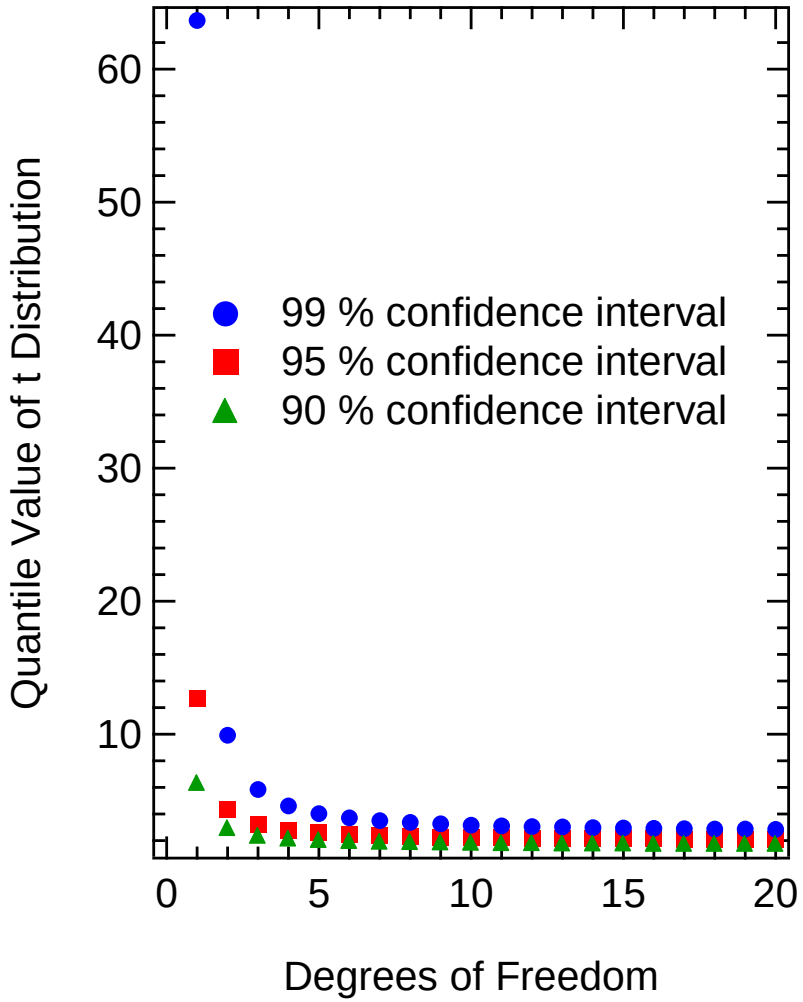


Central Limit Theorem

- **Distribution of means calculated from (an infinite sample of) data from most distributions is approximately normal**
  - **Becomes more accurate with higher number of samples**
  - **Assumes distributions are not peaked close to a boundary and variances are finite**

$$Z = \frac{\bar{x} \pm \mu}{\sigma_x / \sqrt{n}} \Rightarrow \mu = \bar{x} \pm \frac{Z\sigma_x}{\sqrt{n}}$$

Student's t Distribution



- **Used to compute confidence intervals according to**

$$\mu = \bar{x} \pm \frac{st_{\alpha}}{\sqrt{n}}$$

- **Assumes mean and variance estimated by sample values**

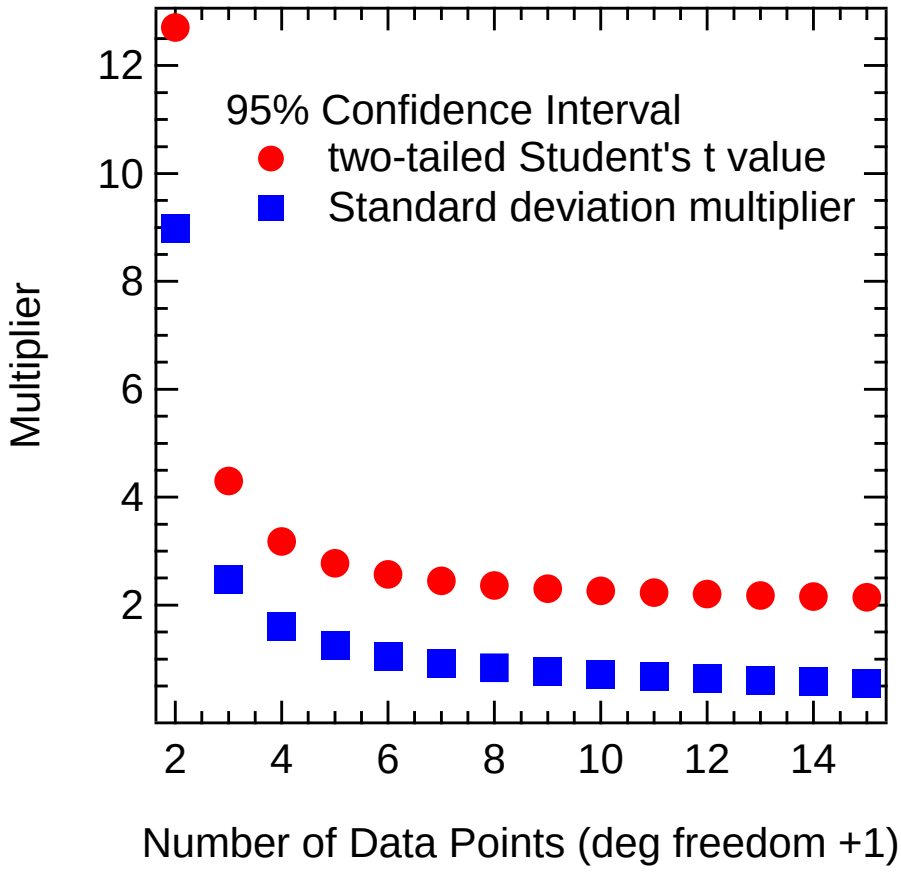
Values of Student's t Distribution

- **Depends on both confidence level being sought and amount of data.**
- **Degrees of freedom generally n-1, with n = number of data points (assumes mean and variance are estimated from data and estimation of population mean only).**
- **This table assumes two-tailed distribution of area.**

	Two-tailed confidence level		
df	90%	95%	99%
1	6.31375	12.7062	63.6567
2	2.91999	4.30265	9.92484
3	2.35336	3.18245	5.84091
4	2.13185	2.77645	4.60409
5	2.01505	2.57058	4.03214
6	1.94318	2.44691	3.70743
7	1.89457	2.36458	3.49892
8	1.85954	2.30598	3.3551
9	1.83311	2.26214	3.24968
10	1.81246	2.22813	3.16918
11	1.79588	2.20098	3.10575
12	1.78229	2.17881	3.0545
13	1.77093	2.16037	3.01225
14	1.76131	2.14479	2.97683
15	1.75305	2.13145	2.9467
16	1.74588	2.1199	2.92077
17	1.73961	2.10982	2.89822
18	1.73406	2.10092	2.87844
19	1.72913	2.09302	2.86093
20	1.72472	2.08596	2.84534
21	1.72074	2.07961	2.83136
22	1.71714	2.07387	2.81875
23	1.71387	2.06866	2.80733
24	1.71088	2.0639	2.79694
25	1.70814	2.05954	2.78743
inf	1.64486	1.95997	2.57583

Sample Size Is Important

- **Confidence interval decreases proportional to inverse of square root of sample size and proportional to decrease in t value.**
- **Limit of t value is normal distribution.**
- **Limit of confidence interval is 0.**



Theory Can Be Taken Too Far

- **Accuracy of instrument ultimately limits confidence interval to something greater than 0.**
  - **Confidence intervals can be smaller than instrument accuracy, but only slightly and if they are you are generally working with poorly designed instruments.**
- **Not all sample means are appropriately treated using Central Limit Theorem and t distribution.**
  - **Computed confidence intervals often include physically unrealizable values when near a boundary, for example, concentrations less than 0 and mole/mass fractions greater than 1.**

Typical Numbers

data points	t quantile	sd multiplier
2	12.7062	8.98464
3	4.30265	2.48414
4	3.18245	1.59122
5	2.77645	1.24166
6	2.57058	1.04944
7	2.44691	0.924846
8	2.36458	0.836004
9	2.30598	0.76866
10	2.26214	0.715353
11	2.22813	0.671807
12	2.20098	0.635368
13	2.17881	0.604293
14	2.16037	0.577382
15	2.14479	0.553781
16	2.13145	0.532862
17	2.1199	0.514152
18	2.10982	0.497288
19	2.10092	0.481984
20	2.09302	0.468014
inf	1.95997	0

- **Two-tailed analysis**
- **Population mean and variance unknown**
- **Estimation of population mean only**
- **Calculated for 95% confidence interval**
- **Based on number of data points, not degrees of freedom**

An Example

- **Five data points with sample mean and standard deviation of 713.6 and 107.8, respectively.**
- **The estimated population mean and 95% confidence interval is:**

$$\mu = \bar{x} \pm \frac{st_{\alpha}}{\sqrt{n}} = 713.6 \pm \frac{107.8 * 2.77645}{\sqrt{5}}$$

$$= 713.6 \pm 133.9$$

$$= 713.6(133.9)$$

Properties of Standard Deviations

$$SD(X) = \sigma_x$$

$$SD(aX) = a\sigma_x$$

$$SD(a + X) = \sigma_x$$

$$SD(XY) \cong \sigma_x\sigma_y - r\bar{x}\bar{y} \neq \sigma_x\sigma_y$$

Point vs. Model Estimation

- **Point estimation**
  - **Characterizes a single, usually global value**
  - **Generally simple mathematics and statistical analysis**
  - **Procedures are unambiguous**
- **Model development**
  - **Characterizes a function of dependent variables**
  - **Complexity of parameter estimation and statistical analysis depend on model complexity**
  - **Parameter estimation and especially statistics somewhat ambiguous**

Overall Approach

- Assume model
- Estimate parameters
- Check residuals for bias or trends
- Estimate parameter confidence intervals
- Consider alternative models

General Confidence Interval

- **Degrees of freedom generally = *n-p*, where *n* is number of data points and *p* is number of parameters**
- **Confidence interval for parameter β given by**

$$\beta_j = \hat{\beta}_j \pm t_{\alpha,n-p}\left(\hat{\sigma}^2 C_{jj}\right)^{1/2}$$

Linear Fit Confidence Interval

- **For intercept:**

$$\beta_0 = \hat{\beta}_0 \pm t_{\alpha,n-2}\left[\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right]^{1/2}$$

- **For slope:**

$$\beta_1 = \hat{\beta}_1 \pm t_{\alpha,n-2}\left(\frac{\hat{\sigma}^2}{S_{xx}}\right)^{1/2}$$

Definition of Terms

$$S_{xx} = \sum_{i}^{n} \left(x_i - \bar{x} \right)^2 =$$

$$\sum_{i}^{n} x_i^2 - \frac{\left(\sum_{i}^{n} x_i \right)^2}{n}$$

Confidence Interval for Y at a Given X

$$\mu_{Y|x} = \hat{\mu}_{Y|x} \pm t_{\alpha,n-2}\left[\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\right]^{1/2}$$

An Example

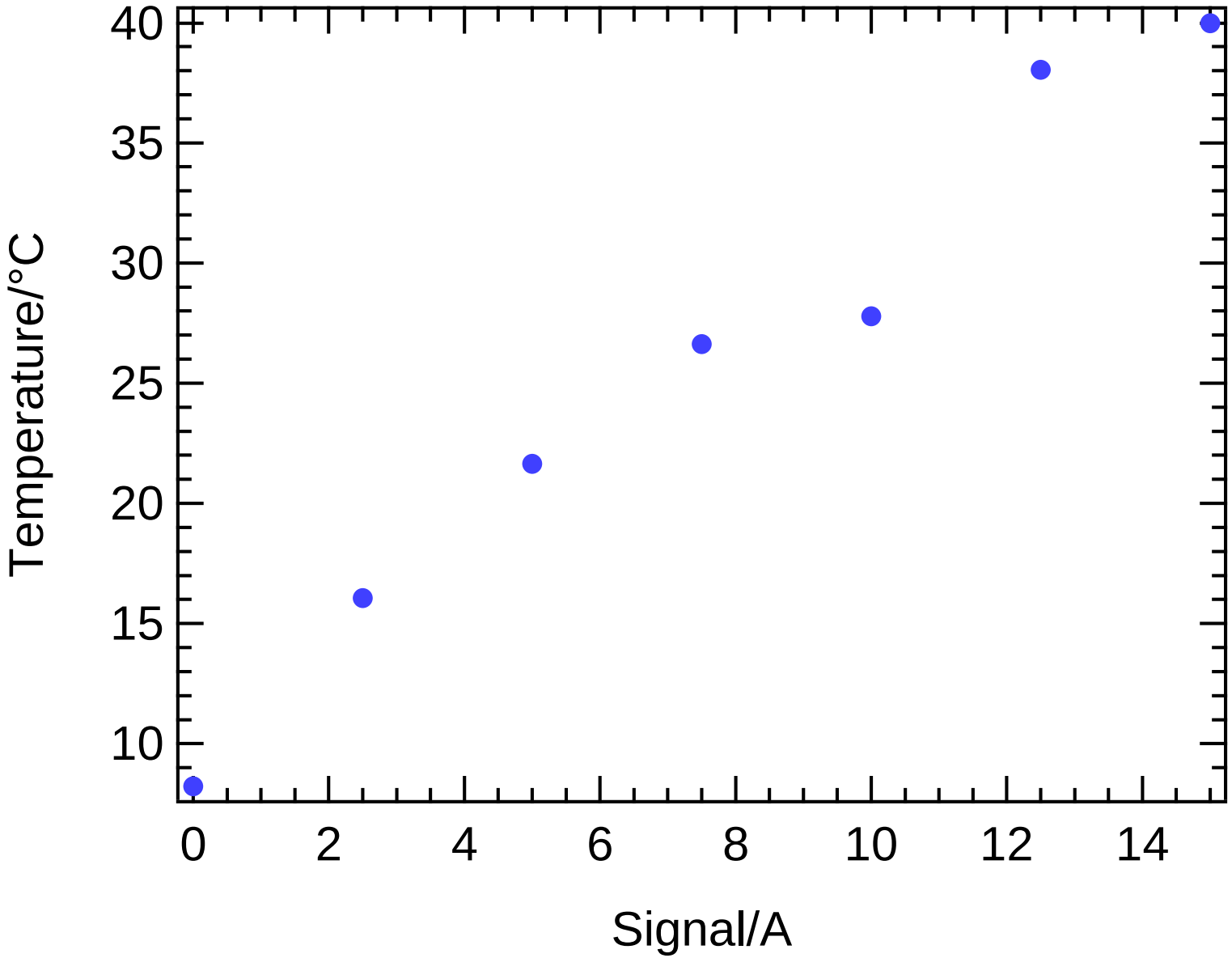
Assume you collect the seven data points shown at the right, which represent the measured relationship between temperature and a signal (current) from a sensor. You want to know how to determine the temperature from the current.

Current/A	Temperature/°C
0	8.22524
2.5	16.0571
5	21.6508
7.5	26.621
10	27.7787
12.5	38.0298
15	39.9741

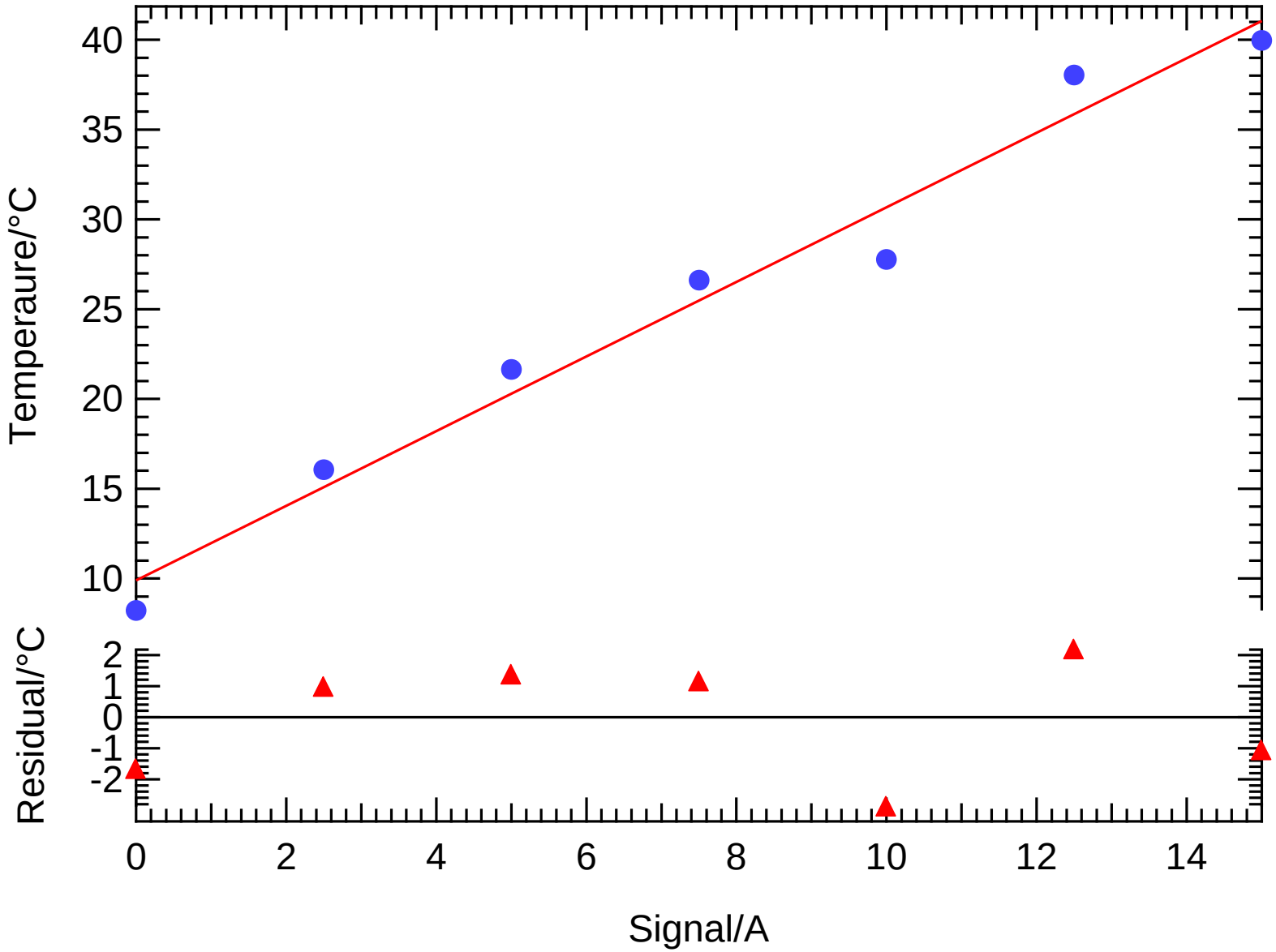
Three Important Things

- **Plot residuals and analyze them for patterns.**
- **Determine model parameters.**
- **Determine confidence intervals for parameters and, if appropriate, for prediction.**

First Plot the Data



Fit Data and Determine Residuals



Determine Model Parameters

$$T = a + b * A$$

$$\hat{a} = 9.91$$

$$\hat{b} = 2.08$$

$$\bar{y}_{resid} = 0.0$$

$$\sigma_{resid} = 1.88$$

$$a = \hat{a} \pm t_{\alpha,n-2}\hat{\sigma}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]^{1/2}$$

$$= 9.91 \pm 3.6$$

$$b = \hat{b} \pm \frac{t_{\alpha,n-2}\hat{\sigma}}{\sqrt{S_{xx}}}$$

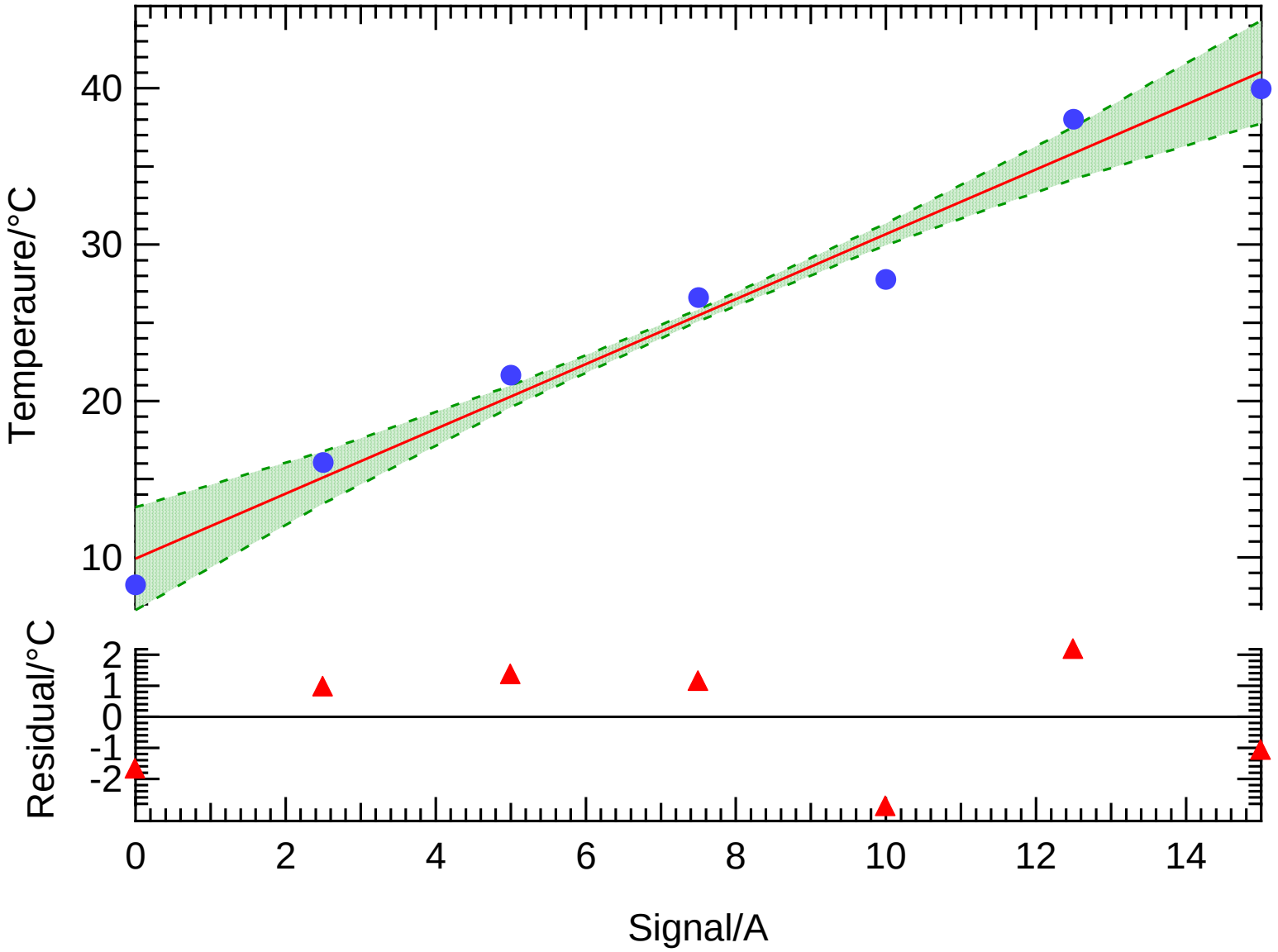
$$= 2.08 \pm 0.399$$

Residuals are easy and accurate means of determining if model is appropriate and of estimating overall variation (standard deviation) of data.

The average of the residuals should always be zero.

These formulas apply only to a linear regression. Similar formulas apply to any polynomial and approximate formulas apply to any equation.

Determine Confidence Interval



Determine Control Points