Practical Basics of Statistical Analysis

David Keffer
Dept. of Materials Science & Engineering
The University of Tennessee
Knoxville, TN 37996-2100
dkeffer@utk.edu
http://clausius.engr.utk.edu/

Governor's School
University of Tennessee, Knoxville
June 8, 2016

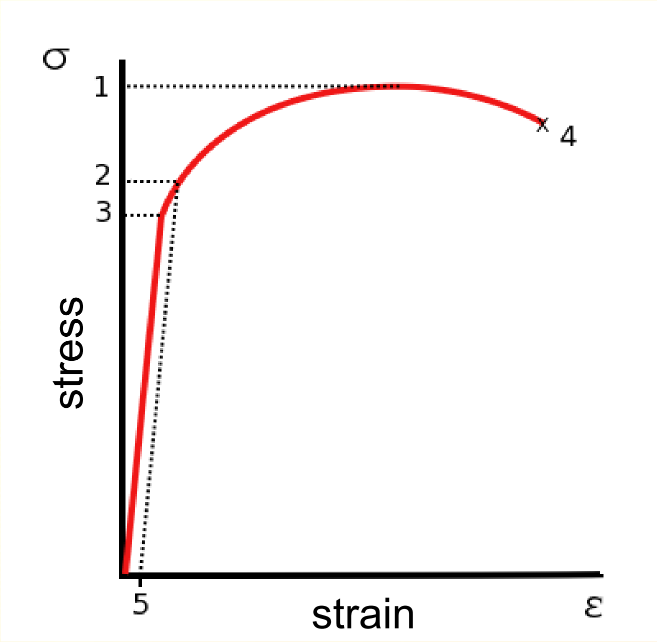
Purpose

Use a materials example to explore basic statistical tools including

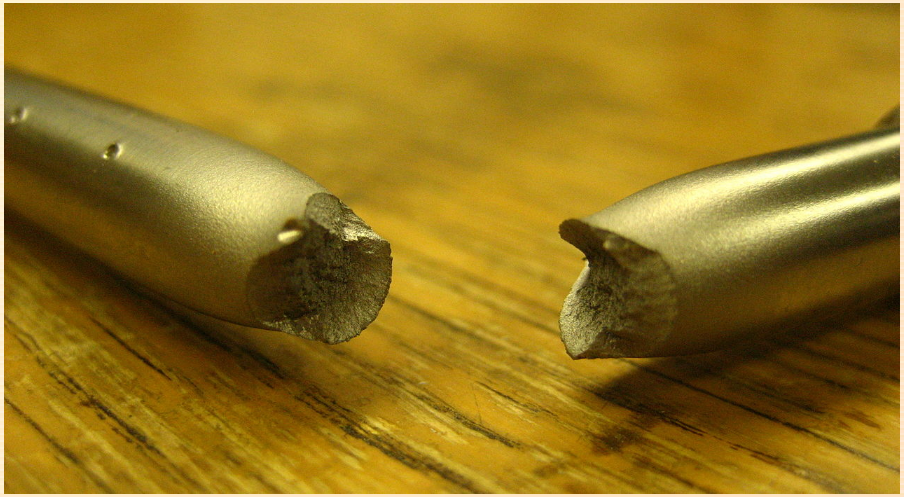
- Mean
- Standard deviation
- Standard error
- Histograms
- Regression

Material Properties have a Probability Distribution

Example: Consider a property like the strain at which fracture occurs in a component.



Stress vs. strain curve typical of aluminum
1. Ultimate tensile strength
2. Yield strength
3. Proportional limit stress
4. Fracture
5. Offset strain (typically 0.2%)

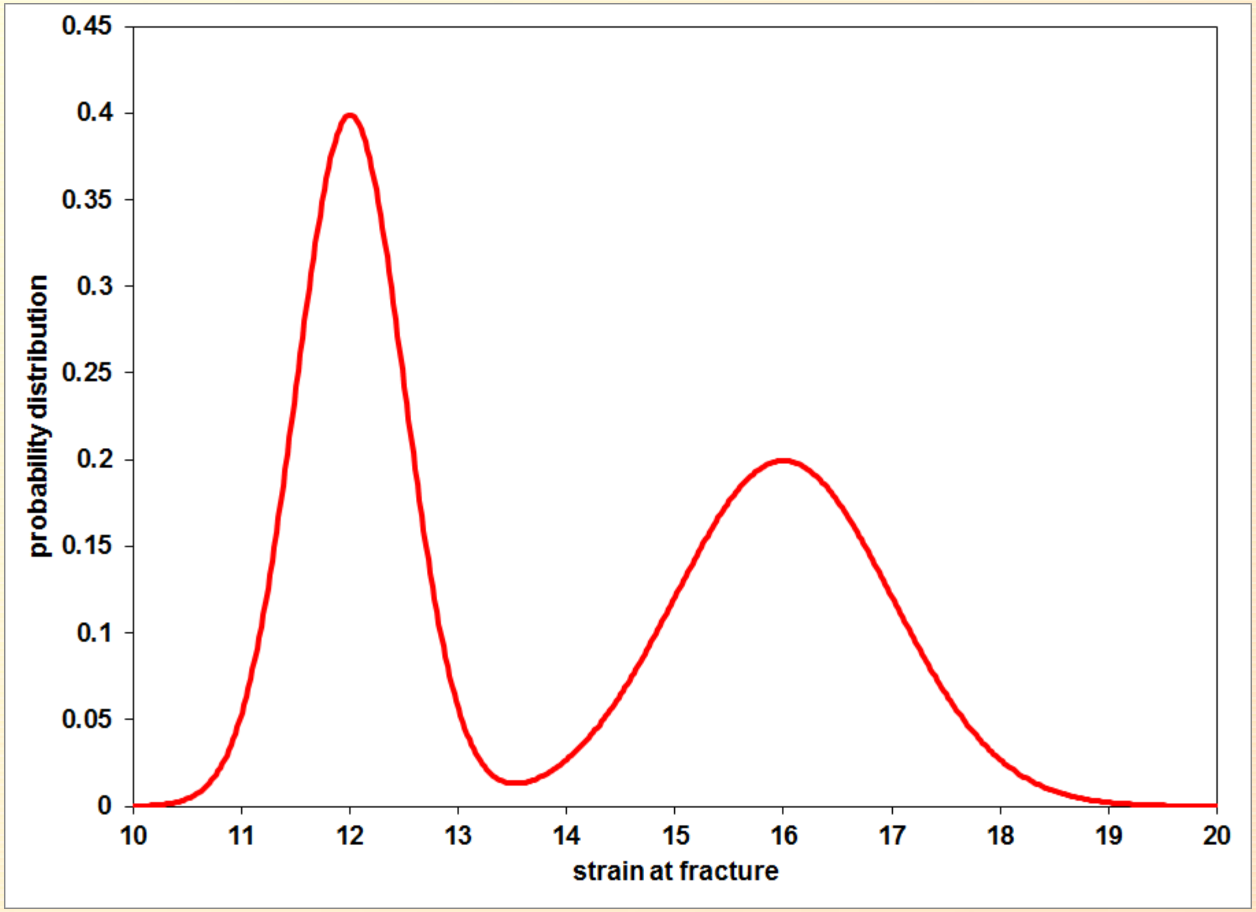


Your Task: Determine the strain at fracture.

http://en.wikipedia.org/wiki/Fracture

Material Properties have a Probability Distribution

What if two shifts use different heat treating procedures resulting in components with two different fracture strains?
The distribution of fracture strains could look something like this:



This “true” probability distribution is unknown! How can you investigate it?

Sampling

Information about the distribution of a property can be obtained from sampling. The "Quality Assurance" (QA) engineer tests a number of components and records the strain at fracture. A mean or average can be evaluated.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

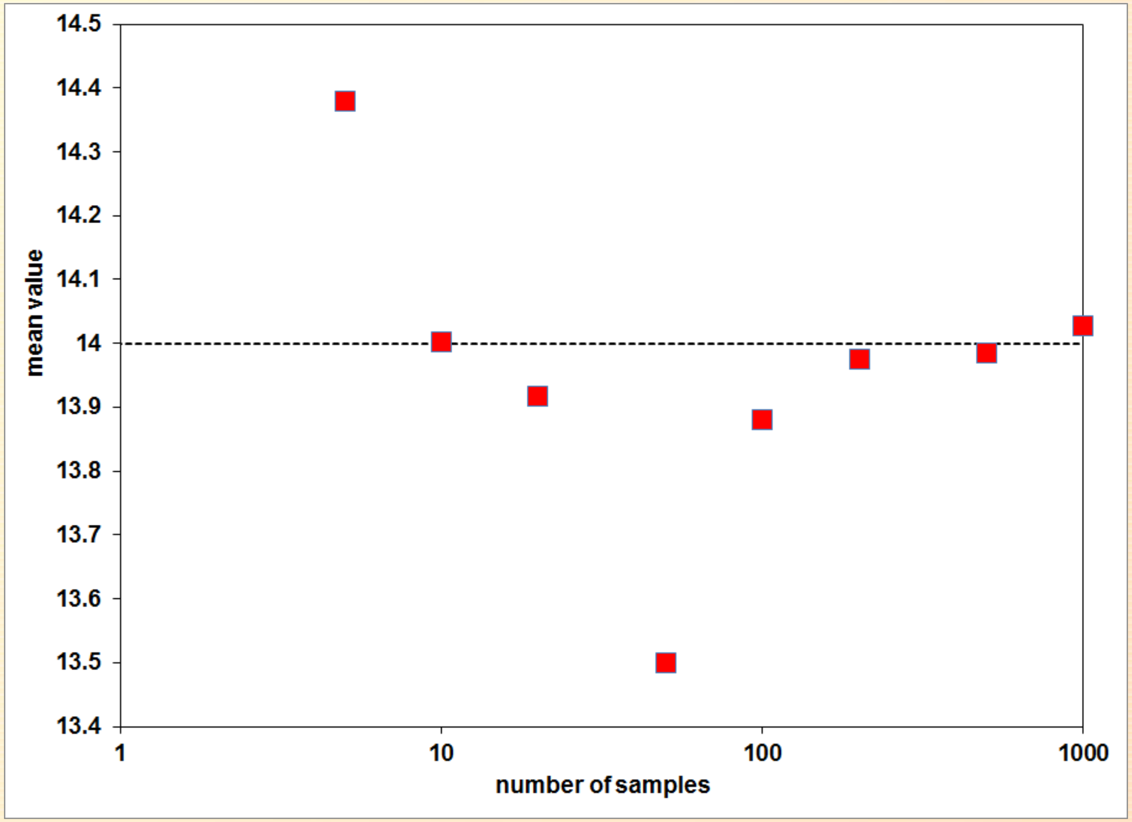
sample #	measurement
1	12.02
2	11.8
3	11.3
4	11.08
5	16.04
6	12.26
7	11.48
8	12.34
9	11.58
10	16.6

n	mean
5	12.42
10	13.286
20	13.519
50	13.8104
100	13.9162
200	13.9768
500	14.02588
1000	14.01478

Mean Value

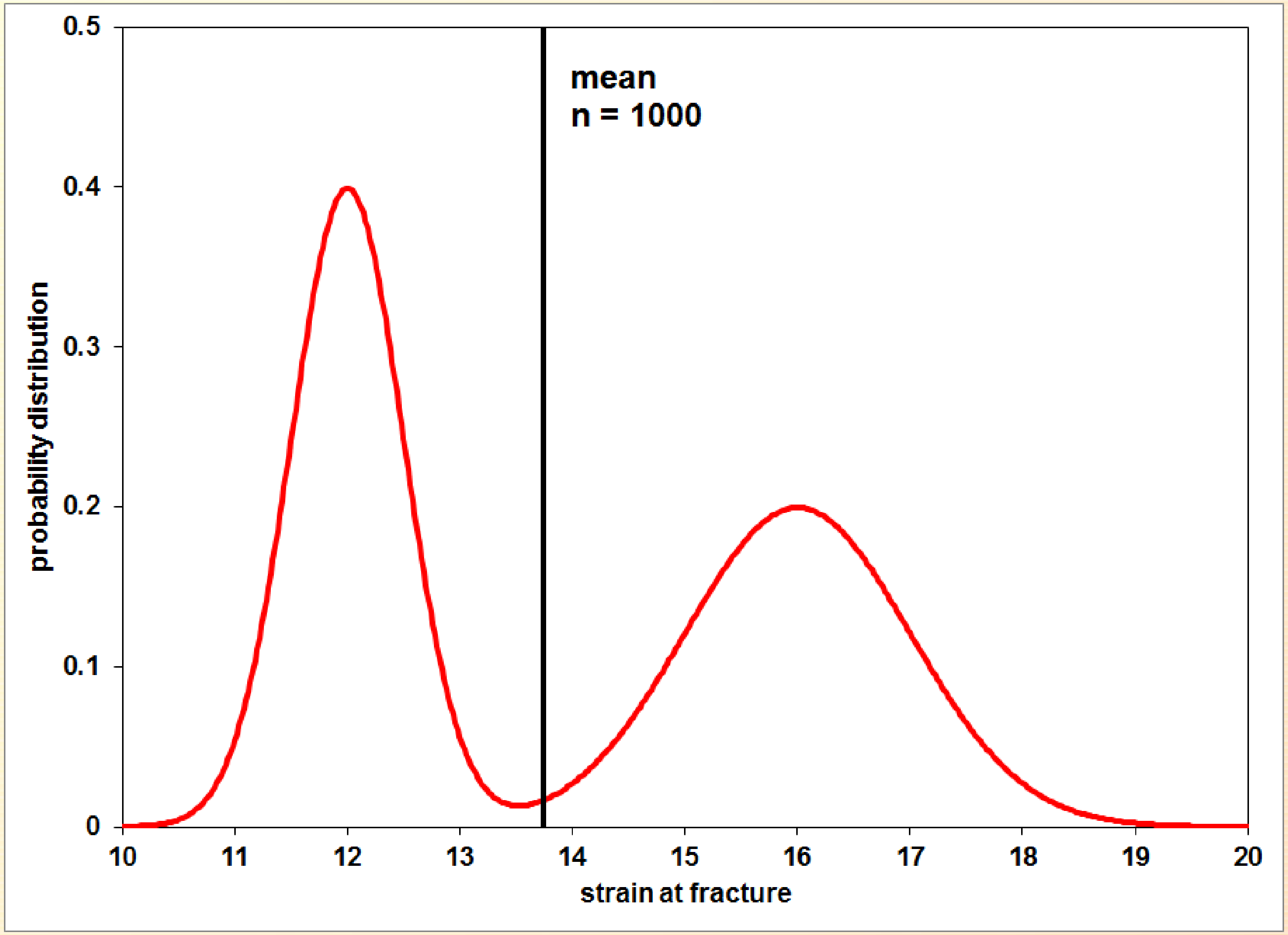
The estimate of the mean gets better with more sample points

n	mean
5	12.42
10	13.286
20	13.519
50	13.8104
100	13.9162
200	13.9768
500	14.02588
1000	14.01478



Mean Value

Even a fairly accurate mean, calculated with a lot of sample points, can't reveal the shape of the underlying distribution.



We might like more information than the mean provides.

Standard Deviation

The standard deviation provides the lowest order description of the distribution of the data around the mean.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

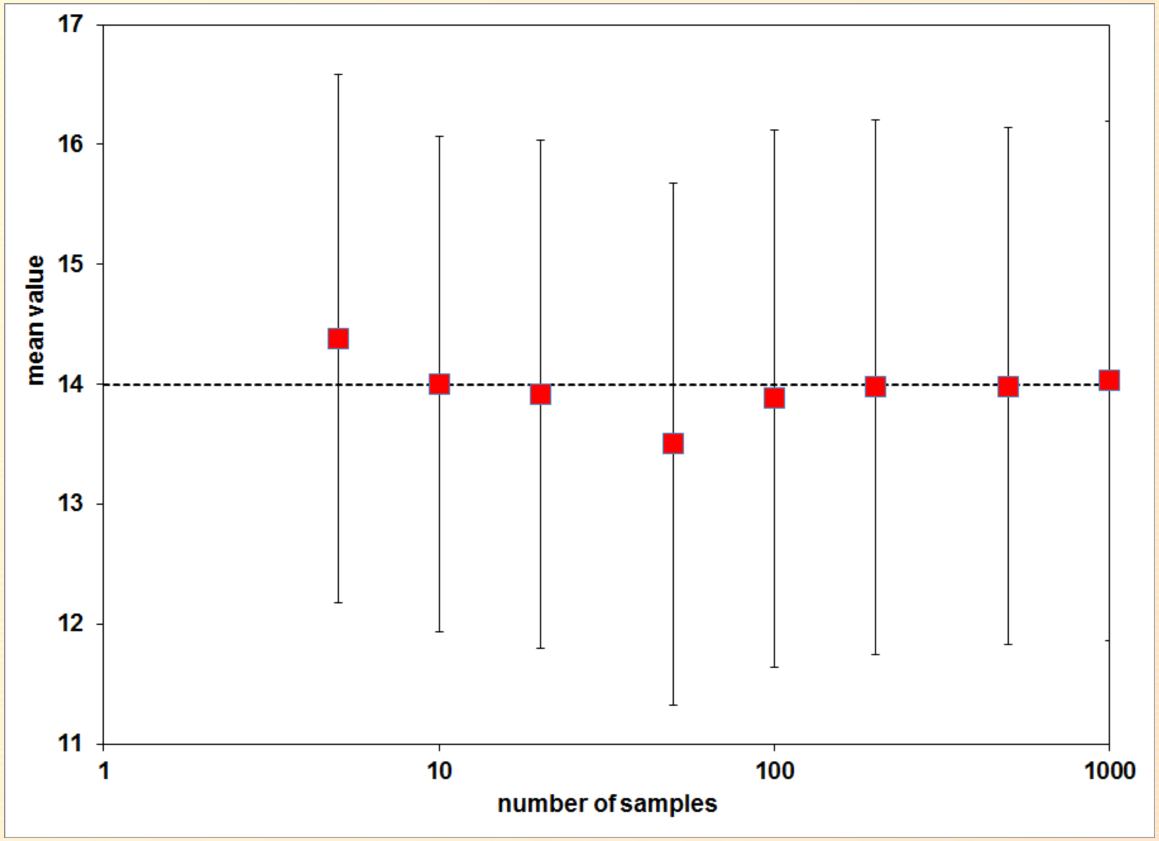
n	mean
5	12.448
10	12.65
20	13.361
50	13.4168
100	13.5132
200	13.6659
500	13.79128
1000	13.71628

n	mean	standard deviation
5	12.42	1.229878
10	13.286	2.231074
20	13.519	2.278351
50	13.8104	2.121264
100	13.9162	2.08423
200	13.9768	2.161243
500	14.02588	2.204589
1000	14.01478	2.177468

Standard Deviation

The estimate of the standard deviation reaches a constant with more sample points.

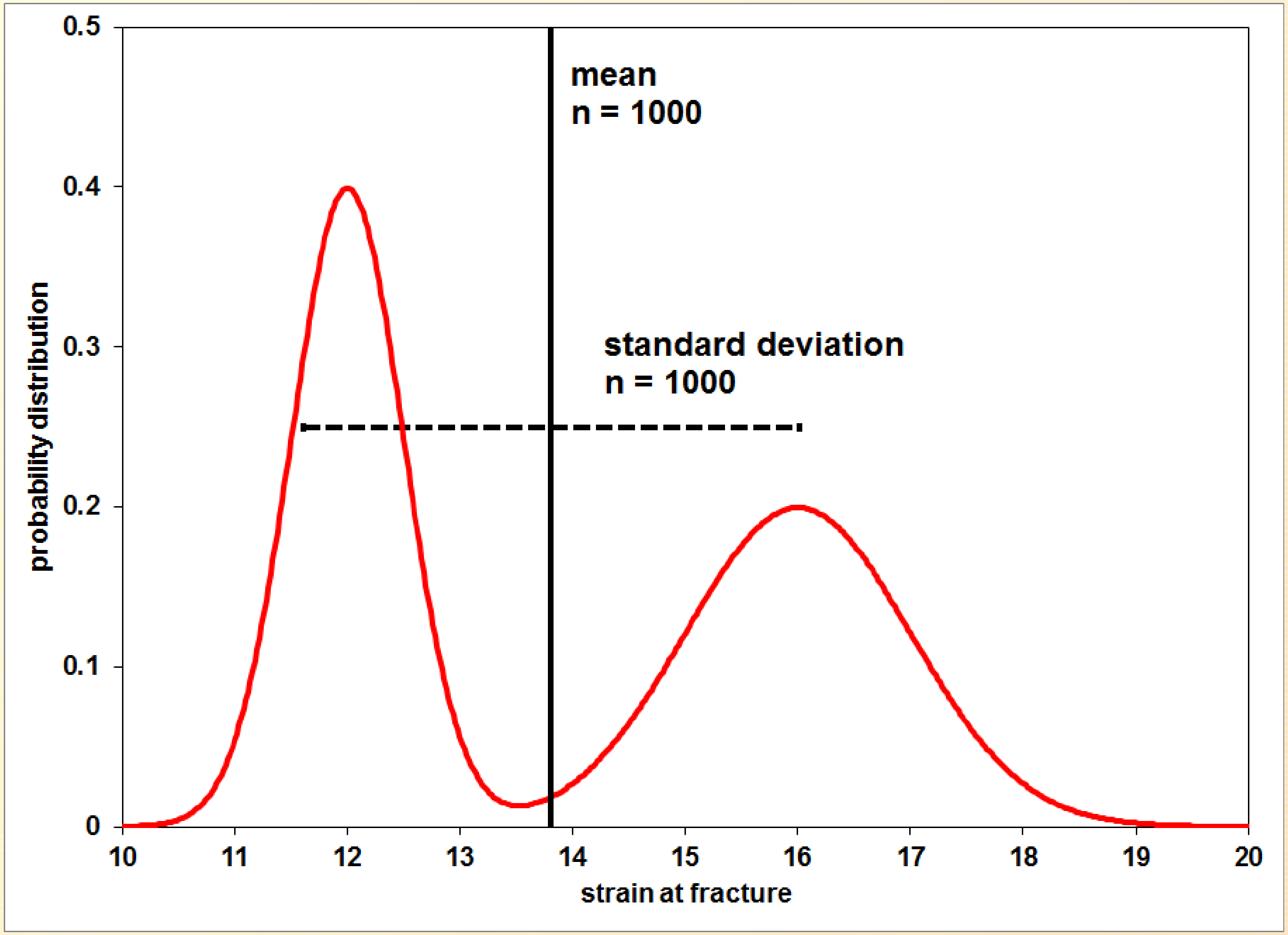
n	mean	standard deviation
5	12.42	1.229878
10	13.286	2.231074
20	13.519	2.278351
50	13.8104	2.121264
100	13.9162	2.08423
200	13.9768	2.161243
500	14.02588	2.204589
1000	14.01478	2.177468



$$\bar{x} \pm s = 14.01 \pm 2.18$$

Standard Deviation

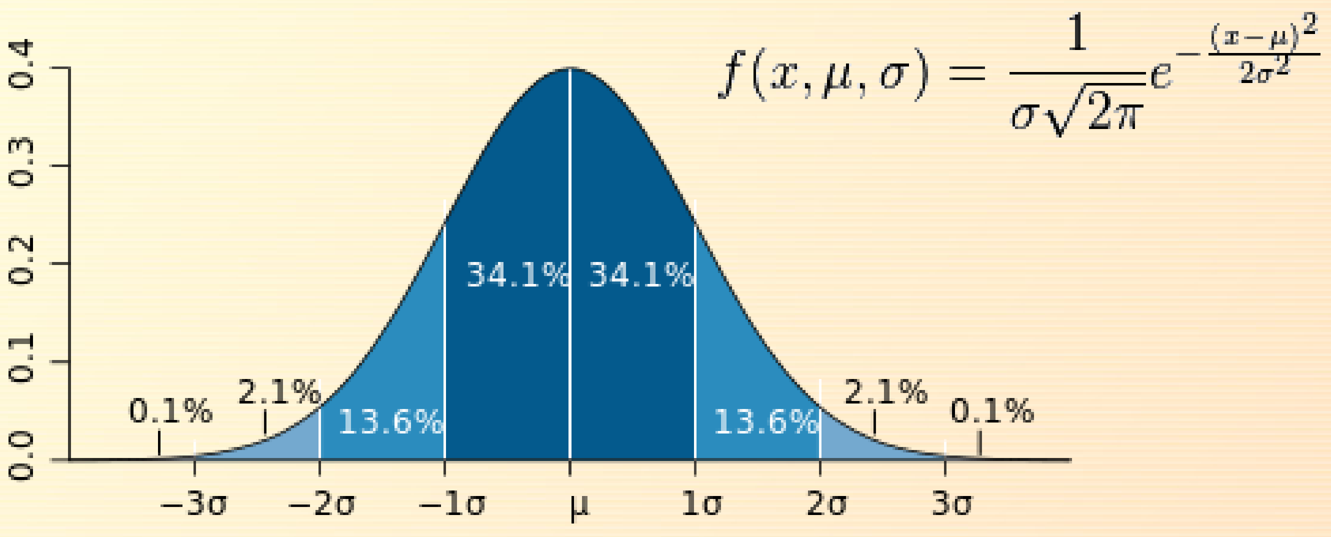
The standard deviation in this example reflects a broad distribution of possible results.



We might like more information than the mean and standard deviation provide.

Distribution of the Sample Mean

The central limit theorem states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed.



normal distribution

http://en.wikipedia.org/wiki/Standard_deviation

http://en.wikipedia.org/wiki/Central_limit_theorem

Standard Error

The standard deviation provides a description of the distribution of the data around the mean.

$$SE = \frac{s}{\sqrt{n}}$$

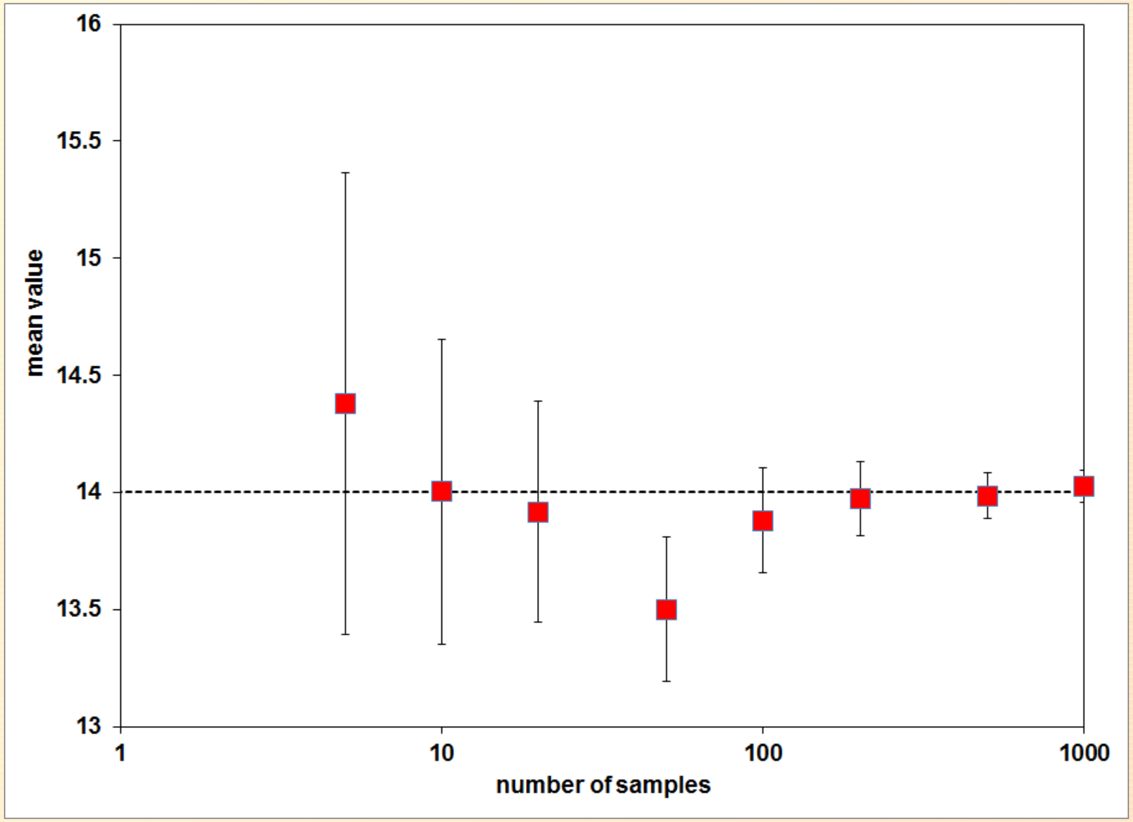
n	mean
5	12.448
10	12.65
20	13.361
50	13.4168
100	13.5132
200	13.6659
500	13.79128
1000	13.71628

n	mean	standard deviation	standard error
5	12.42	1.229878	0.550018
10	13.286	2.231074	0.705528
20	13.519	2.278351	0.509455
50	13.8104	2.121264	0.299992
100	13.9162	2.08423	0.208423
200	13.9768	2.161243	0.152823
500	14.02588	2.204589	0.098592
1000	14.01478	2.177468	0.068858

Standard Error

The standard error is a measure of uncertainty in the sample mean.
The standard error becomes smaller with more sample points

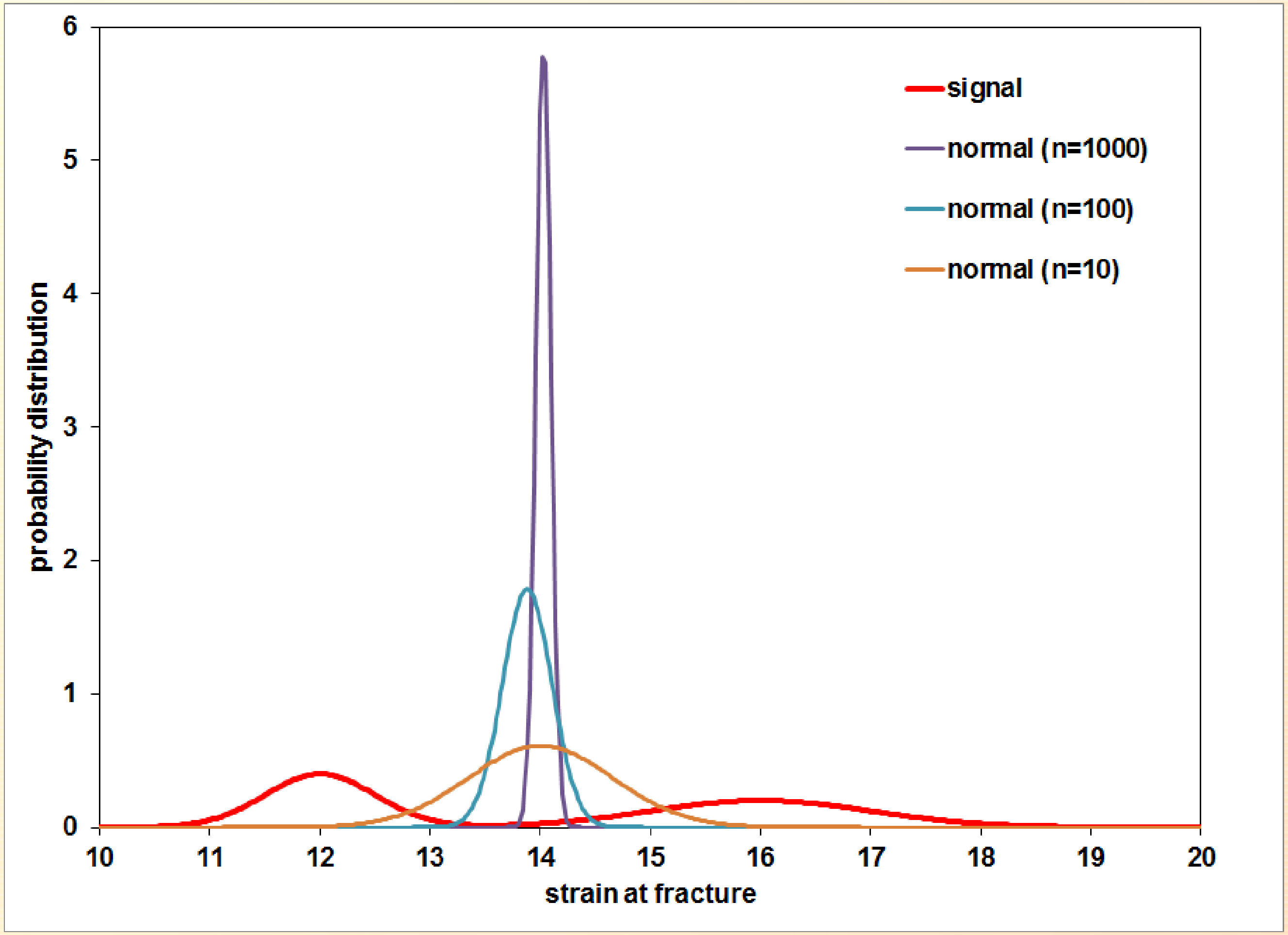
n	mean	standard error
5	12.42	0.550018
10	13.286	0.705528
20	13.519	0.509455
50	13.8104	0.299992
100	13.9162	0.208423
200	13.9768	0.152823
500	14.02588	0.098592
1000	14.01478	0.068858



$$\bar{x} \pm SE = 14.01 \pm 0.07$$

Standard Error

The standard error represents your uncertainty in the sample mean, but does not tell you much about the actual distribution..

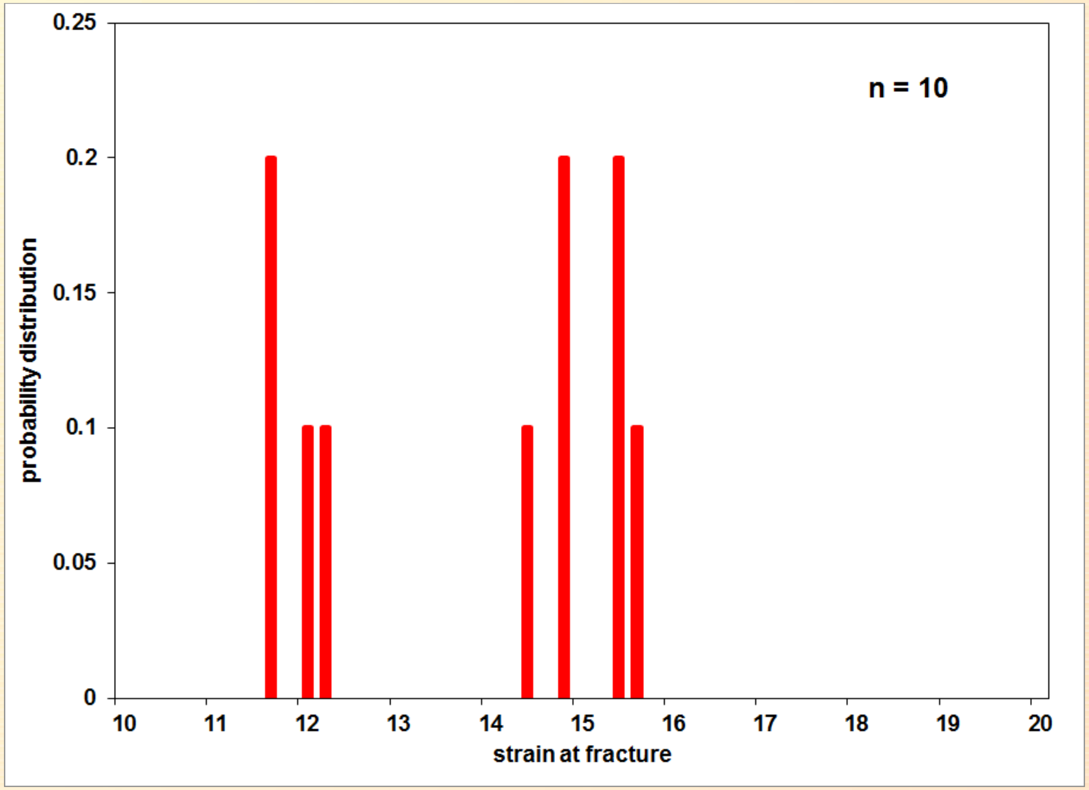


We might like more information than the mean and standard error provide.

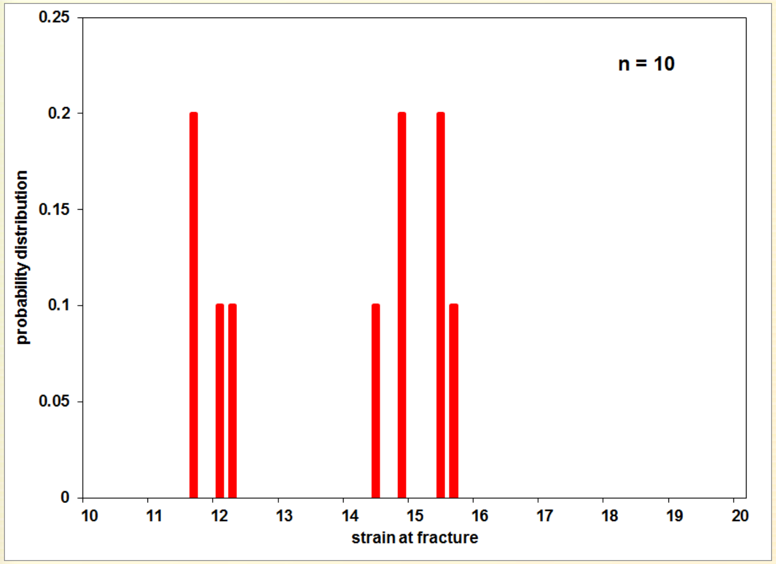
Histograms

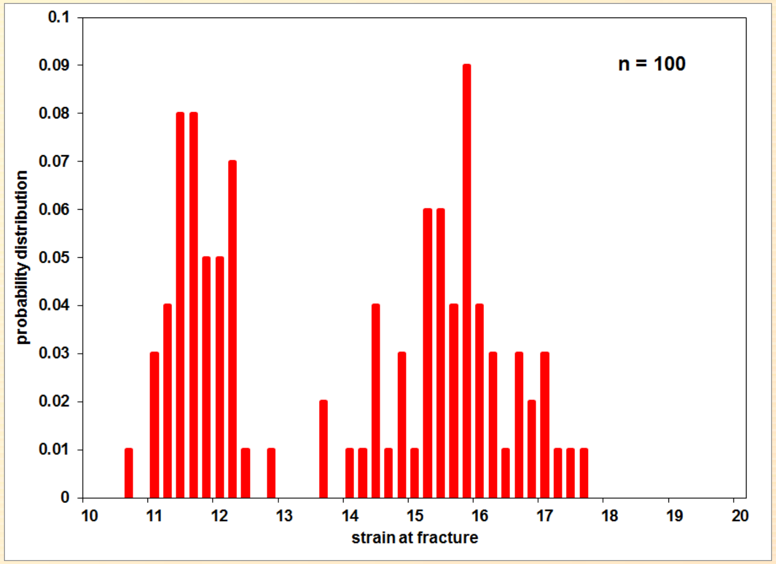
Information about the distribution of a property can be obtained from sampling. The “Quality Assurance” (QA) engineer tests a number of components and records the strain at fracture. A histogram can be created.

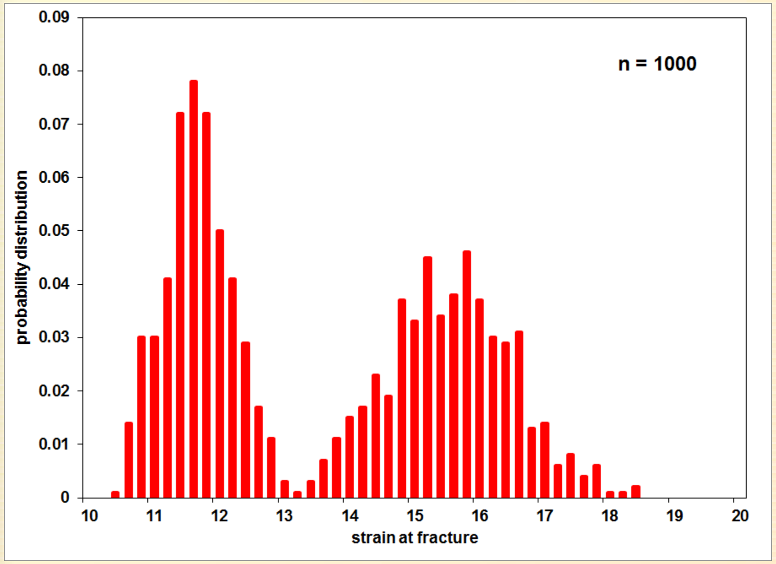
sample #	measurement
1	15.44475
2	16.26772
3	11.98035
4	12.00618
5	16.22892
6	12.02725
7	12.64226
8	11.7119
9	15.63842
10	16.15753

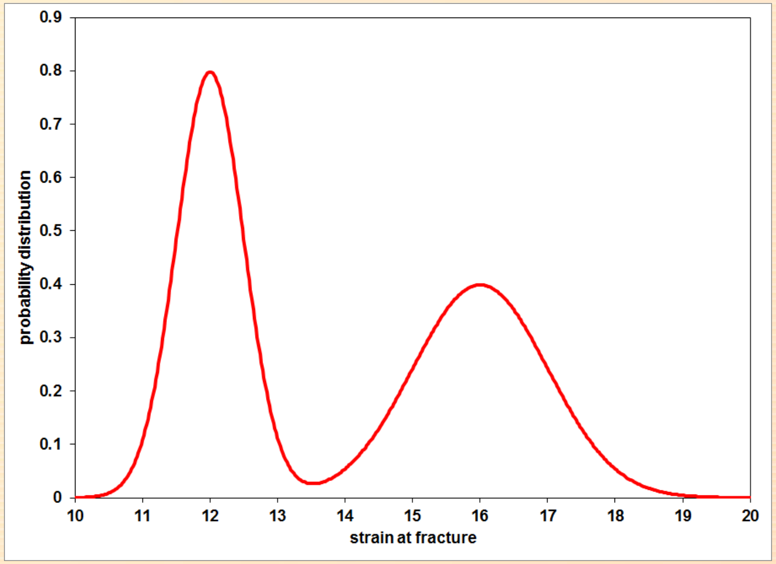


Histograms become more accurate with more sampling









Regression

A linear regression provides the coefficients for a linear model relating a dependent and independent variable.

$$y = mx + b$$

Consider the strain at fracture for a series of components in which the heat treatment time is varied.

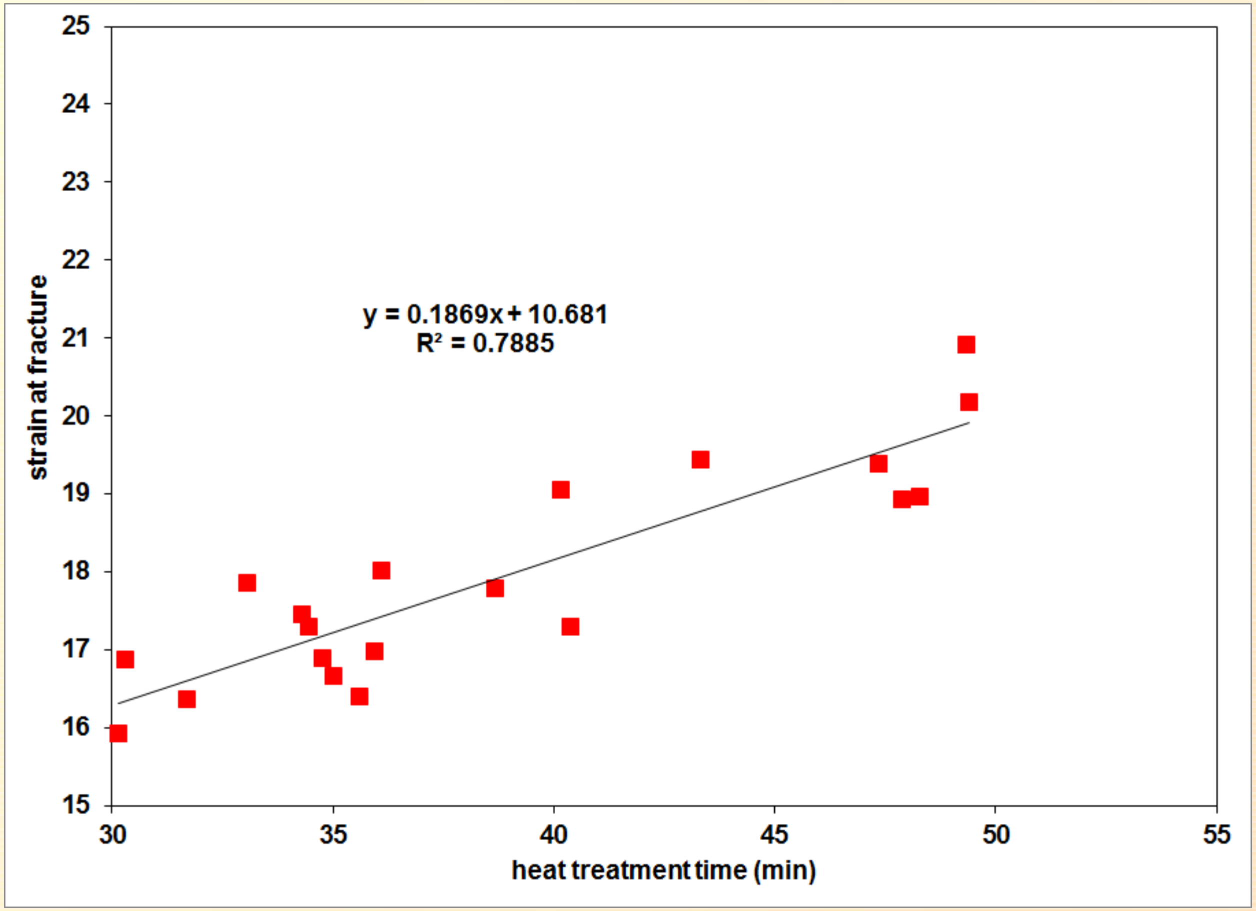
$$\varepsilon_f = mt + b$$

If we can find the missing coefficients (slope and intercept) then we can use them to predict the strain at fracture for a given heat treatment time.

sample	treatment time	strain at fracture
1	35.60675	16.39818
2	40.16145	19.06037
3	38.65641	17.78325
4	40.37866	17.2992
5	35.95295	16.97849
6	48.27652	18.96844
7	36.10229	18.02034
8	49.39935	20.17243
9	47.34459	19.39355
10	43.31971	19.4442
11	33.06743	17.85545
12	49.33015	20.9129
13	47.87165	18.93009
14	34.77224	16.89122
15	34.99737	16.65559
16	34.296	17.45434
17	34.43596	17.28692
18	30.14439	15.92876
19	30.30892	16.87964
20	31.70115	16.37002

Regression

Frequently, the results of a regression are presented as a plot.



The R^2 Measure of Fit is bound between 0 (no fit) and 1 (perfect fit).

Document Access

These slides and a sample excel file presenting examples for

- Mean
- Standard deviation
- Standard error
- Histograms
- Regression

are located online at

http://utkstair.org/clausius/docs/governorsschool/index.html