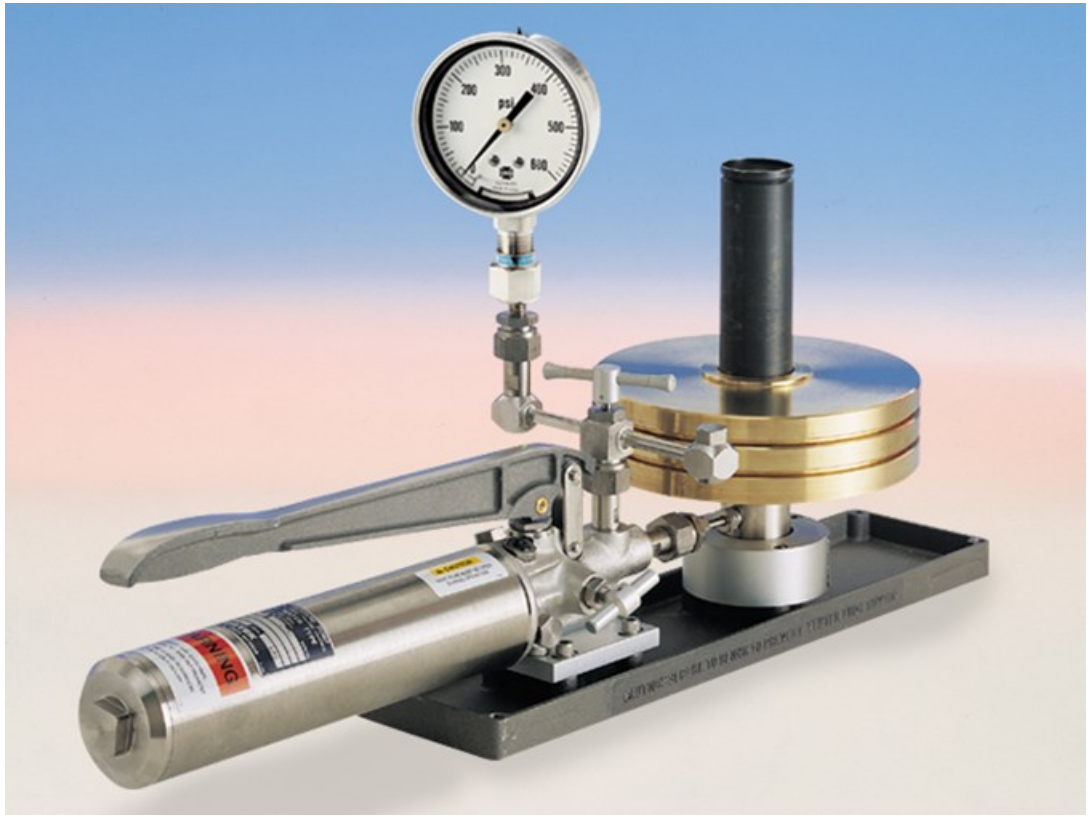
Lecture 2
Statistics, probability density functions, linear regression, confidence intervals

The material in this lecture covers Ch. 1.4 and 4

Consider a dead weight pressure test to calibrate a pressure gage

- Primary standard
- Load with weight through piston. Pressure is mg/A where A is the area of the piston and m is mass of the standard weight.
- Measure gauge reading for several different dead weights to construct a calibration curve.
- Determine hysteresis by increasing then decreasing mass.



Trial #	Po (kPa)
1	10
2	10.19
3	10.26
4	10.19
5	10.21
6	10.12
7	9.94
8	10.11
9	10.07
10	9.87
11	10.03
12	10.16
13	10.43
14	10.2
15	10.22
16	10.09
17	9.95
18	10.08
19	10.02
20	9.77

Assume our dead weight allows us to apply a "true" pressure of P_i = 10.000 ± 0.001 kPa.

Measure the pressure from the gauge 20 times. We calculate the mean and standard deviation using the following formulae.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

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

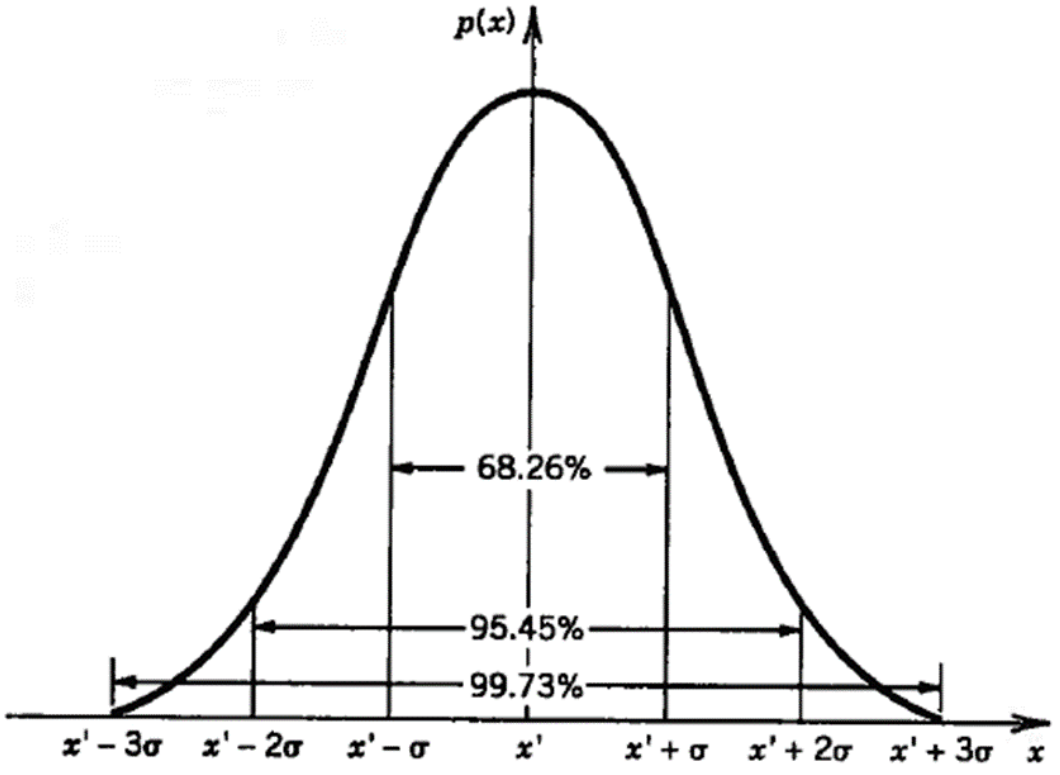
For the data in the table the mean pressure and the standard deviation are:

$$\bar{P}_o = 10.11 kPa \qquad s_{P_o} = 0.14 kPa$$

All this tells us is that our limited data set has a mean of 10.11 kPa and a standard deviation of 0.14 kPa. It does not allow us to predict anything about the next measurement or what confidence interval we can assign to a single measurement or the average of the next group of measurements.

Infinite statistics, Normal Probability Density Function (PDF)

- If we measure something an infinite number of times, the results will cluster about the mean with a distribution of results about the mean, x′.
- By normalizing the results by the number of measurements, we get the fraction of results at a given measurement value which is equal to the fractional probability that that value will occur.
- The plot of probability, p(x), vs. measurement value(x- x′) is called the probability density function (PDF). There is a list of useful PDF's shown in Table 4.2, page 123-124
- The one we typically use is called the Normal or Gaussian PDF. The breadth of the distribution is characterized by σ, the standard deviation. It is informative to observe what fraction of results are expected to be within a certain multiple of σ from the mean.

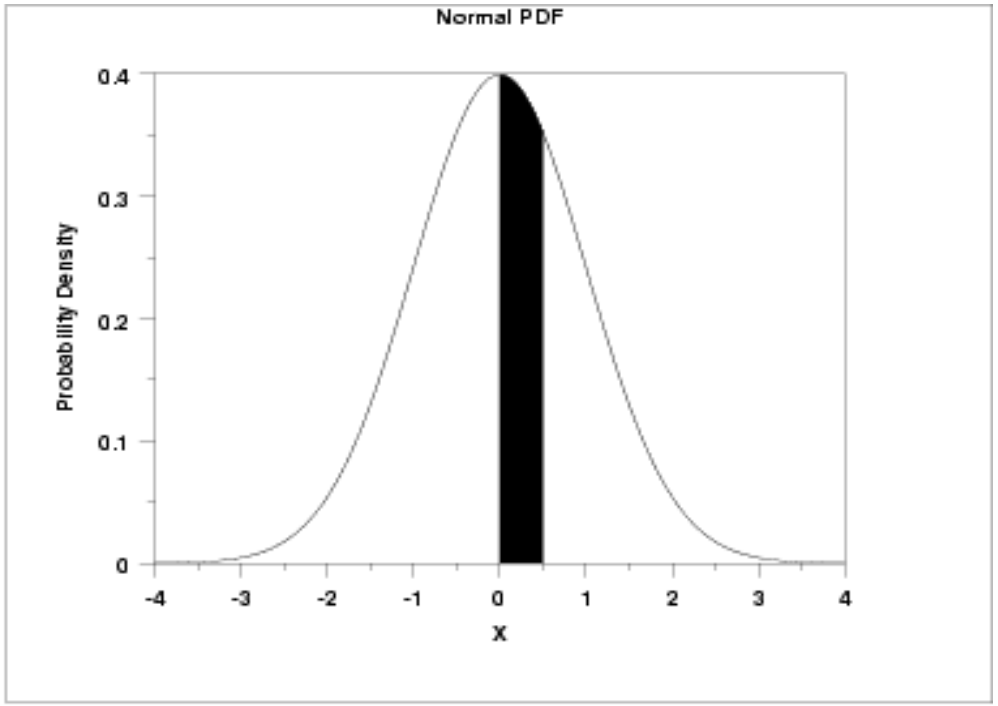


$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x - x')^2}{\sigma^2} \right]$$

How do we use PDF's

- We typically want to predict the likelihood that the measured value will fall within some sort of ***confidence interval***. Looking at the Normal PDF on the previous page, it shows that the likelihood of the measured value being within one standard deviation of the mean is P= 0.6826. We would say the expected value of $x = x' \pm \sigma$ (68.26%)
- The integral of a PDF from $-\infty$ to ∞ is equal to 1.
- We want the cumulative distribution. That is obtained by integrating the PDF between specified limits. There are several ways to do this.
  - Use of tables (explained on next few slides)
  - Matlab function cdf
  - Excel function NORM.DIST

The normal PDF is symmetrical. We can determine P for a given confidence interval by integrating p(x) between limits. It is useful to normalize the range by the standard deviation, σ. We can determine P for a given interval, ± z_1, by multiplying the integral from 0 to z_1 by two.



http://www.itl.nist.gov/div898/handbook/eda/section3/gif/norpdftb.gif

$$P\left(x' - \delta x \le x \le x' + \delta x\right) = \int_{x'-\delta x}^{x'+\delta x} p(x)dx$$

$$\beta = \frac{x - x'}{\sigma}; \quad \mathrm{dx} = \sigma \, \mathrm{d}\, \beta$$

$$z_1 = \frac{x_1 - x'}{\sigma}$$

$$P\left(-z_1 \le \beta \le z_1\right) = \frac{1}{\sqrt{2\pi}} \int_{-z_1}^{z_1} \mathrm{e}^{-\beta^2/2} d\beta = 2\left(\frac{1}{\sqrt{2\pi}} \int_{0}^{z_1} \mathrm{e}^{-\beta^2/2} d\beta\right)$$

For example, suppose we want to determine the probability that our measurement will like between ±1.37σ. We identify P/2 by adding the left hand vertical column for 1.3 with the top horizontal row for 0.07 to get 1.37. The value of P/2 is in the table (circled). The probability that the value will lie between ±1.37σ is 0.4147 x 2 = 0.8294.

We can read the table to find the interval corresponding to a certain probability. For instance, if we want to know the interval for a probability of 0.99, we look on the table for 0.495 and read the table value for z_1. This would be at ±2.575σ by interpolation.

If we want to find the likelihood that a measurement will be < x′+ 1.8σ, we use the knowledge that the value below zero is 0.5 and add 0.5 + 0.4641 = 0.9641. Similar arguments for the measurement to be < x′- 1.8σ yield 0.5 – 0.4641 = 0.0359.

Table 4.3 Probability Values for Normal Error Function: One-Sided Integral Solutions for $p(z_1) = \frac{1}{(2\pi)^{1/2}} \int_0^{z_1} e^{-\beta^2/2} d\beta$

$z_1 = \frac{x_1 - x'}{\sigma}$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1809	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2794	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4292	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4758	0.4761	0.4767
2.0	0.4772	0.4778	0.4803	0.4788	0.4793	0.4799	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.49865	0.4987	0.4987	0.4988	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990

$p(x)$

Matlab Cumulative Distribution Function

y=cdf('Normal',x,xbar,sdev)

xbar = mean
sdev=standard deviation

Returns probability of data being between - ∞ and x

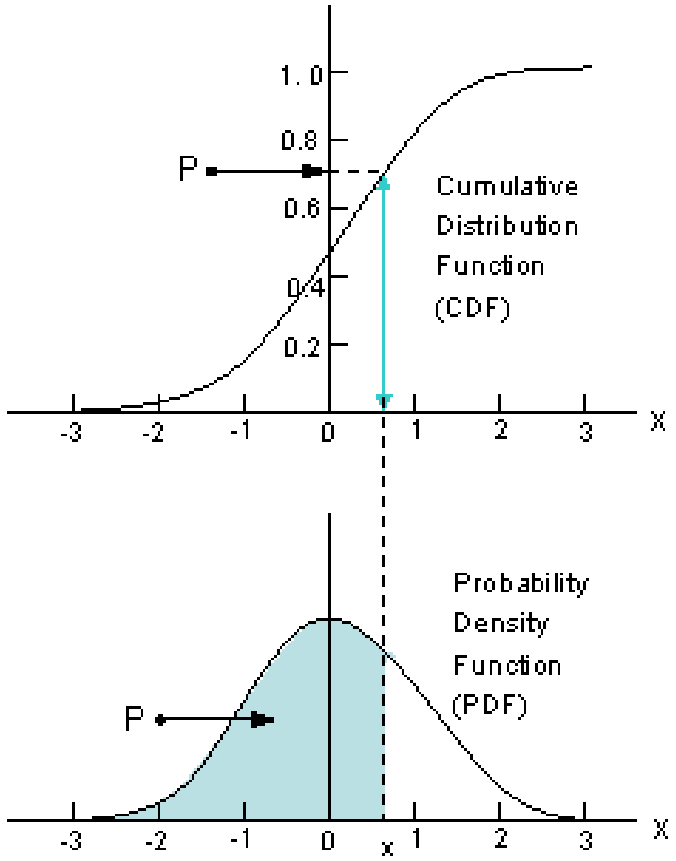
yrange=y=cdf('Normal',x1,xbar,sdev)-cdf('Normal',x2,xbar,sdev)

Returns the probability that the values will lie between x2 and x1 where x2 < x1

```
y=cdf('Normal',7, 5, 2)- cdf('Normal',3,5,2);
```

= 0.6827

Cumulative distribution functions



Relations Between Two Different Typical Representations of a Population

http://home.ubalt.edu/ntsbarsh/business-stat/CdfAndPdf.gif

Excel CDF

NORM.DIST(x,mean,standard_dev,true) returns the cumulative distribution from -∞ up to a value of x

Statistics of finite sized data sets: Student's t distribution

Small data sets (samples) do not always accurately estimate the properties of the infinite population. The t distribution increases the size of the confidence interval to attain a desired probability.

$$\bar{x} = sample\ mean = \frac{1}{N}\sum_{i=1}^{n} x_i$$

$$s_x = sample\ standard\ deviation = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

$$x_i = \bar{x} \pm t_{\nu,P} s_x \ (P\%)$$

$$t = \frac{\bar{x} - x'}{s_x / \sqrt{N}};\ \nu = N - 1\ (degrees\ of\ freedom)$$

For a sample size N = 20 and a likelihood of 95%, $t_{19,95} = 2.093$

This means the measured values have a 95% likelihood of falling between $\pm 2.093 s_x$ of the sample mean.

Table 4.4 Student's t Distribution

ν	t_{50}	t_{90}	t_{95}	t_{99}
1	1.000	6.314	12.706	63.657
2	0.816	2.920	4.303	9.925
3	0.765	2.353	3.182	5.841
4	0.741	2.132	2.770	4.604
5	0.727	2.015	2.571	4.032
6	0.718	1.943	2.447	3.707
7	0.711	1.895	2.365	3.499
8	0.706	1.860	2.306	3.355
9	0.703	1.833	2.262	3.250
10	0.700	1.812	2.228	3.169
11	0.697	1.796	2.201	3.106
12	0.695	1.782	2.179	3.055
13	0.694	1.771	2.160	3.012
14	0.692	1.761	2.145	2.977
15	0.691	1.753	2.131	2.947
16	0.690	1.746	2.120	2.921
17	0.689	1.740	2.110	2.898
18	0.688	1.734	2.101	2.878
19	0.688	1.729	2.093	2.861
20	0.687	1.725	2.086	2.845
21	0.686	1.721	2.080	2.831
30	0.683	1.697	2.042	2.750
40	0.681	1.684	2.021	2.704
50	0.680	1.679	2.010	2.679
60	0.679	1.671	2.000	2.660
∞	0.674	1.645	1.960	2.576

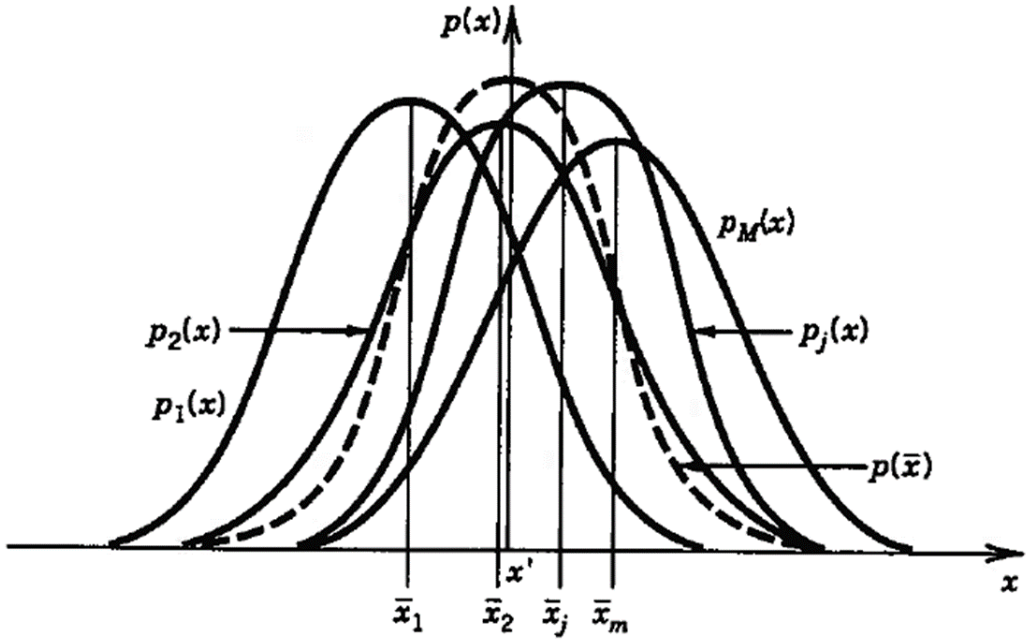
We often need to take small samples of a population to estimate the statistical properties. The sample mean is not the same as the population mean as shown in the diagram to the right. The dashed line is the infinite population PDF. The solid lines are the PDF for smaller samples and all have different sample means.

We can define a *standard deviation of the mean* as;

$$s_{\bar{x}} = \frac{s_x}{\sqrt{N}}$$

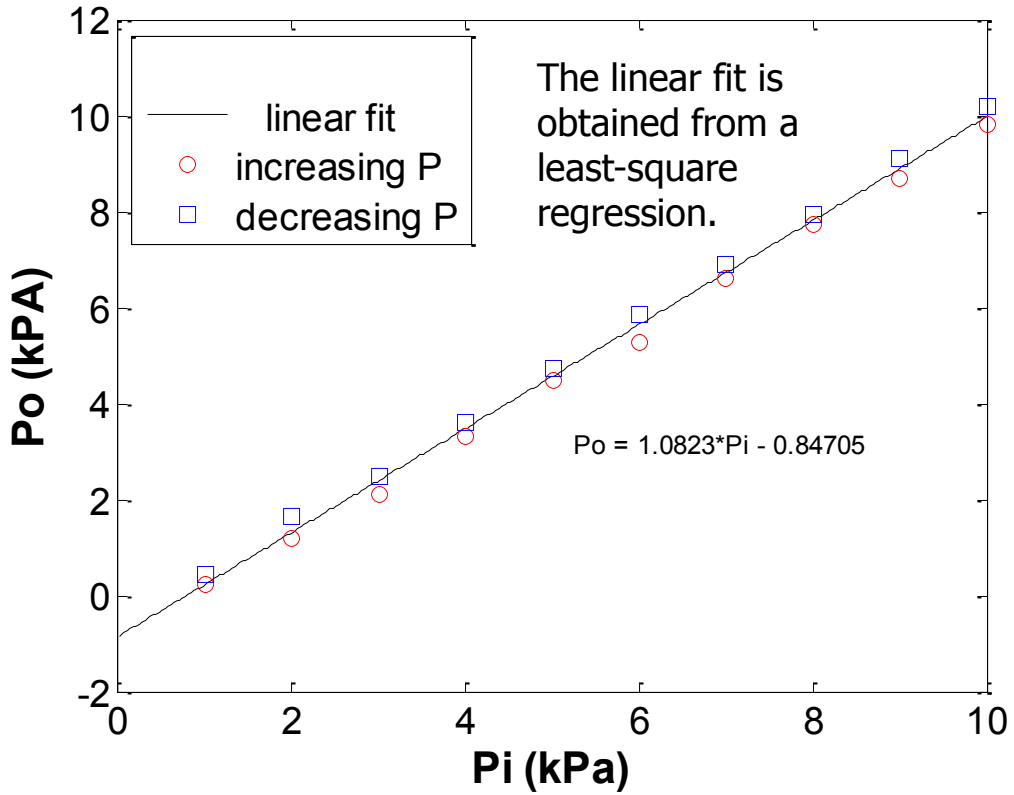
This means the true (infinite population) mean, x′, is related to the sample mean by;

$$x' = \bar{x} \pm t_{\nu,P} s_{\bar{x}}$$



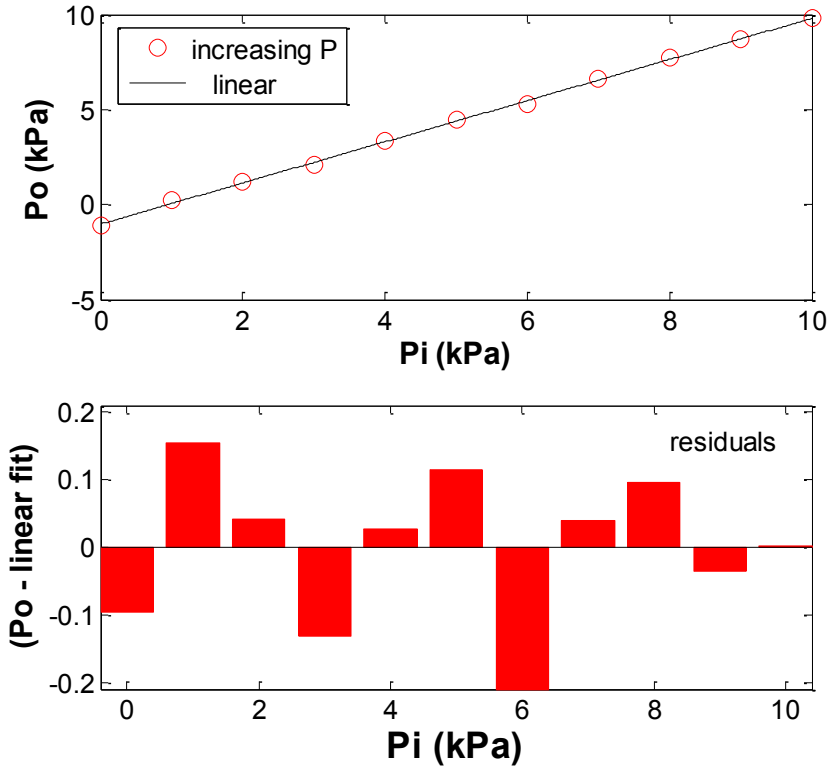
Back to the calibration of the pressure gauge…..We need to know the sensitivity and the confidence interval for the P_o - P_i relationship. We obtain data over the expected range of the transducer first increasing P and then decreasing to detect hysteresis. The data set looks linear and we fit the entire data set to a first order relationship using linear regression.

Pi (kPa)	Po (kPa)
0.000	-1.12
1.000	0.21
2.000	1.18
3.000	2.09
4.000	3.33
5.000	4.50
6.000	5.26
7.000	6.59
8.000	7.73
9.000	8.68
10.000	9.80
10.000	10.20
9.000	9.10
8.000	7.92
7.000	6.89
6.000	5.87
5.000	4.71
4.000	3.62
3.000	2.48
2.000	1.65
1.000	0.42
0.000	-0.69



The **sensitivity** of the instrument is defined as the slope of the calibration curve. For a linear curve it is a constant, while for a nonlinear curve it varies (i.e., it is a function of the input variable).

We used linear regression to determine the line that best describes the data. One of the ways to determine whether the "model" for the data describes the data well. If the residuals, the difference between the data and the prediction are random over the range, the model is considered good. If you observe residuals of one sign in the middle of the range and the opposite sign near the ends of the measured range, the fit is not considered to be as good.



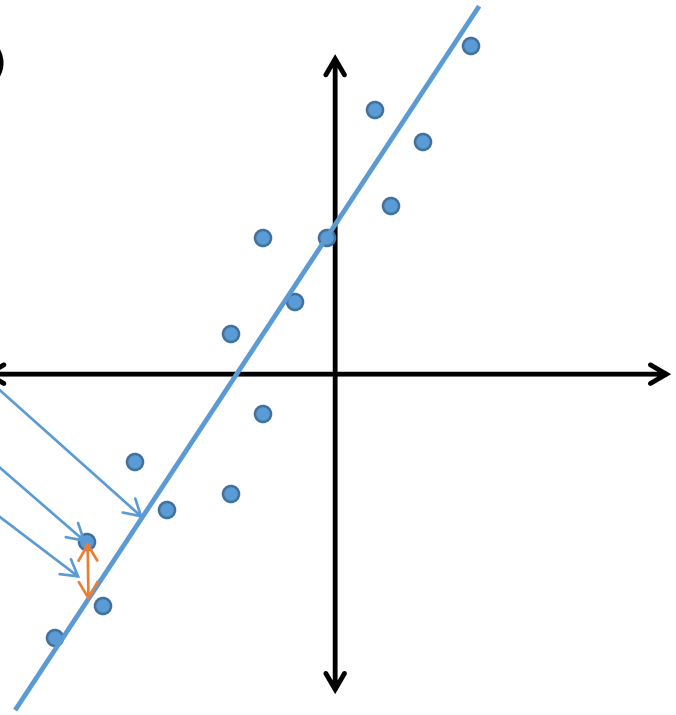
The following slides will describe how to quantify how "good" the measurements are. We will start with basic statistics for single quantity measurements and then move to quantifying the quality of a fit.

Linear regression: Least squares fit

Definition of error between expected values (the best fit line) and the actual values

observed expected

$$E_i = Y_i - y_i = Y_i - (mx_i + b)$$



The error for each point, E_i is defined as which is the difference between the observed values, Y_i , and the expected value, y_i , which is described by the equation of the line.

We square the error and sum all of the individual squared errors to obtain the total squared error for N data points.

$$S = \sum_{i=1}^{N} \left(Y_i - mx_i - b \right)^2$$

$$S = \sum_{i=1}^{N} \left(Y^2{}_i - 2mx_i Y_i - 2bY_i + 2mbx_i + m^2 x_i^2 + b^2 \right)$$

$$S = \sum_{i=1}^{N} Y^2{}_i - 2m \sum_{i=1}^{N} x_i Y_i - 2b \sum_{i=1}^{N} Y_i + 2mb \sum_{i=1}^{N} x_i + m^2 \sum_{i=1}^{N} x_i^2 + Nb^2$$

If we take the derivative of the total squared error with respect to the coefficients of the best fit line, m and b, and set both derivatives to zero, we can find the values of m and b which give the minimum squared error. (When you take the derivatives, remember that the individual summations are just constants as far as the derivative is concerned.

$$\frac{dS}{db} = -2\sum_{i=1}^{N} Y_i + 2m\sum_{i=1}^{N} x_i + 2Nb = 0$$

$$\frac{dS}{dm} = -2\sum_{i=1}^{N} x_i Y_i + 2b\sum_{i=1}^{N} x_i + 2m\sum_{i=1}^{N} x_i^2 = 0$$

Since we have two equations and two unknowns, we can explicitly solve for m and b.

$$m = \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} Y_i - N\sum_{i=1}^{N} x_i Y_i}{\left(\left(\sum_{i=1}^{N} x_i\right)^2 - N\sum_{i=1}^{N}\left(x_i\right)^2\right)} \qquad b = \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i Y_i - \sum_{i=1}^{N} Y_i \sum_{i=1}^{N}\left(x_i\right)^2}{\left(\left(\sum_{i=1}^{N} x_i\right)^2 - N\sum_{i=1}^{N}\left(x_i\right)^2\right)}$$

The polyfit command in Matlab provides the coefficients for up to a ninth order polynomial. For the equation below, the command:

p = polyfit(x,y,3)

will return the following values to the array, p, for arrays x and y that contain a series of values (that must exceed the order of the polynomial):

p(1) = a_3
p(2) = a_2
p(3) = a_1
p(4) = a_o

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 \, \mathrm{x}^3 .....$$

See the file Multiple Regression in Excel.xlsx in Course Information (will go over in class, too)

How do we establish confidence intervals for the fit? We define the *standard error of the fit* as;

$$y_c = a_o + a_1 x + a_2 x^2 + \ldots a_m x^m$$

$$s_{yx} = \sqrt{\frac{\sum_{i=1}^{N} \left(y_i - y_{ci} \right)^2}{\nu}}$$

$$\nu = N - (m+1)$$

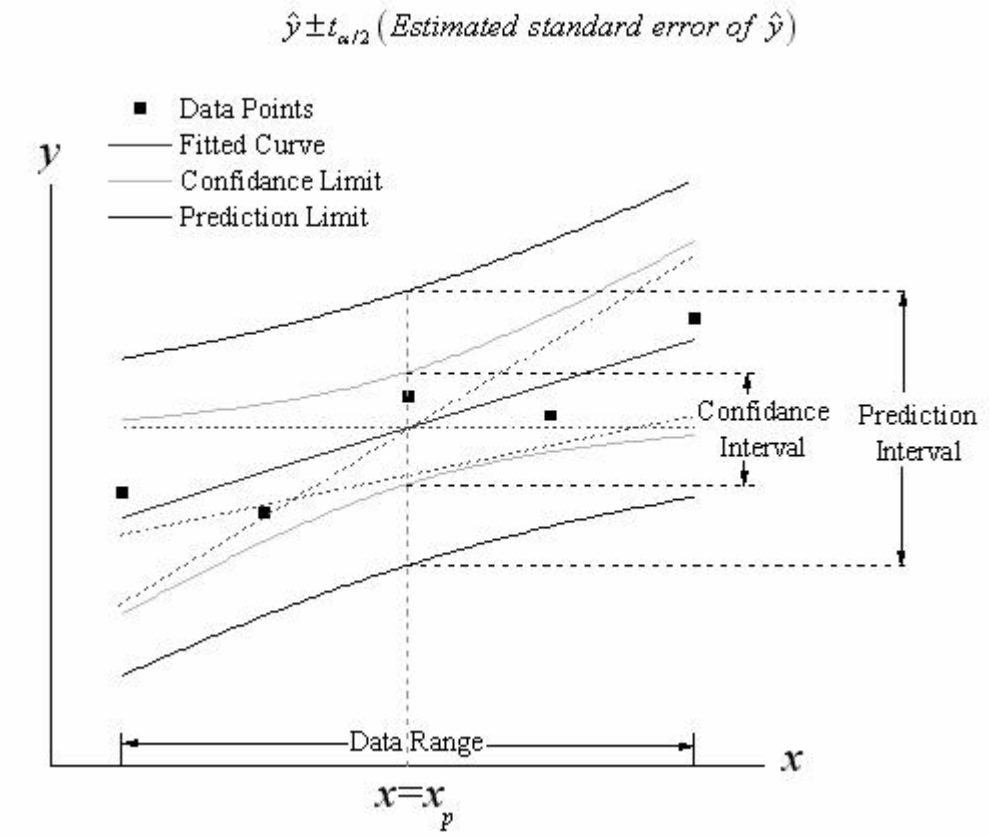
"m" is the order of the fit. Most of the time we will use first order (linear) fits.

The confidence interval *of the fit* is a function of the independent variable (x). The confidence interval is smallest at the middle of the measurement range and highest near the extremes. It is defined as;

$$y(x) = y_c(x) \pm t_{\nu,P} s_{yx} \left[\frac{1}{N} + \frac{\left(x - \bar{x} \right)^2}{\sum_{i=1}^{N} \left(x_i - \bar{x} \right)^2} \right]^{1/2} \qquad (P\%)$$

The 95% **confidence bands** enclose the area that you can be 95% sure contains the true curve. It gives you a visual sense of how well your data define the best-fit curve. It is closely related to the 95% **prediction bands** , which enclose the area that you expect to enclose 95% of future data points. This includes both the uncertainty in the true position of the curve (enclosed by the confidence bands), and also accounts for scatter of data around the curve. Therefore, prediction bands are always wider than confidence bands.

https://www.youtube.com/watch?v=o0UESA3UZss



Precision in estimating the slope and intercept of the fit (Equations 4.42 and 4.43)

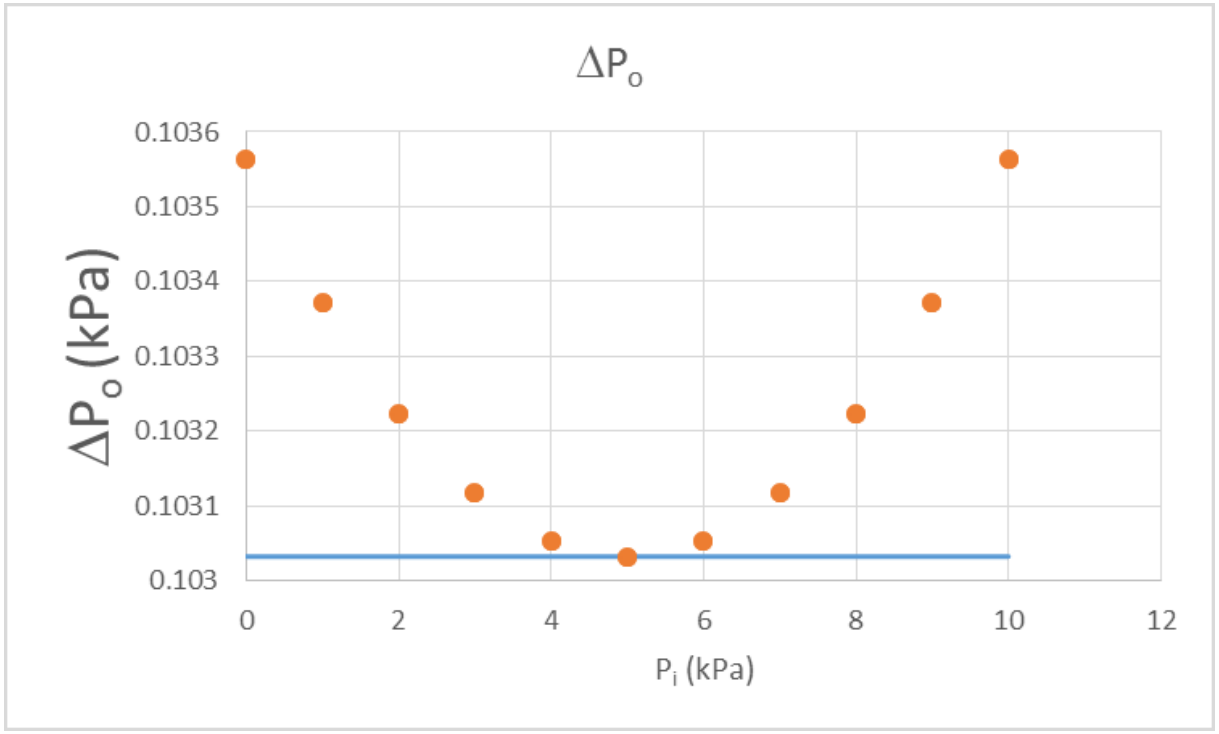
$$y = a_o + a_1 x$$

$$s_{a1} = \frac{s_{yx}}{\sqrt{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}}$$

$$s_{a0} = s_{yx} \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}}$$

Going back to the pressure calibration data, we want to determine the ΔP_o for the fit at a 95% confidence level. Since m = 1, ν = 20 - (1 + 1) = 18. Therefore, $t_{\nu,P}$ = 2.101. The value of s_{yx} = 0.219. The table below shows the ΔP_o calculated using the formula from the previous slide. Note that ΔP_o is not a strong function of P_i.

The blue line is an approximation for the confidence interval; $t_{\nu,P} s_{yx} / \sqrt{N}$



Pi	Po	ΔPo
0	-1.12	0.10356336
1	0.21	0.10337258
2	1.18	0.10322395
3	2.09	0.10311765
4	3.33	0.10305382
5	4.5	0.10303254
6	5.26	0.10305382
7	6.59	0.10311765
8	7.73	0.10322395
9	8.68	0.10337258
10	9.8	0.10356336
10	10.2	0.10356336
9	9.1	0.10337258
8	7.92	0.10322395
7	6.89	0.10311765
6	5.87	0.10305382
5	4.71	0.10303254
4	3.62	0.10305382
3	2.48	0.10311765
2	1.65	0.10322395
1	0.42	0.10337258
0	-0.69	0.10356336

The confidence interval of the fit is different from the confidence interval for the measurement in the equation below. This is the confidence interval for any measurement. We will evaluate this in the second lab. The plot below shows the confidence interval for the pressure gauge calibration data.

$$y(x) = y_c(x) \pm t_{\nu,P} s_{yx} \left[1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right]^{1/2} \qquad (P\%)$$