Measurements to Statistics

Collection and Interpretation of Data

Part II

Sample mean vs. population mean

Given the sample mean what is the (elusive) population mean, μ?

Measurand – particular quantity (unknown parameter) subject to measurement.

Confidence Interval

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter.

The confidence intervals of a measurand is calculated using a statistical tool; Student's t. t – test.

The range of the confidence interval gives a range of plausible values for the unknown parameter and therefore how uncertain we are about the unknown parameter (precision).

Confidence intervals are convenient and informative than the hypothesis tests results where decisions are made either to "reject H_0" (H_0 = null hypothesis) or "don't reject H_0".

Confidence intervals are calculated so that one can be confident at a confidence level (a percentage) the range in which the parameter value would be; like 90%, 95%(norm), 99%, 99.9% (or whatever the % is desired) of the range of values of unknown parameter.

Hypothesis Testing and the Statistics t-Test

The t-test is probably the most commonly used in Statistical Data Analysis procedure for hypothesis testing.

There are several kinds of t-tests, but the most common is the "two-sample t-test" also known as the "Student's t-test" or the "independent samples t-test".

The two sample t-test simply tests whether or not two independent populations have different mean values on some measure.

The null hypothesis, assumes no difference between the two populations to be true until proven wrong; i.e. there no difference between these two populations.
Just because averages of two data sets are different, it does not necessarily mean that the data sets are different.

The statistics t-test allows us to answer this question;

"Are the two groups have different average scores represent a real difference between the two populations, or just a chance difference in our samples?"

by using the t-test statistic to determine a *p-value* that indicates how likely we could have gotten these results by chance.

By convention, if there is *a less than 5%* probability of getting the observed differences by chance, (i.e. greater than 95% probability the difference is not by chance) *we reject the null hypothesis and say we found a statistically significant difference between the two groups.*

Two major uses of Student's t:

1. Statistical Evaluation of the mean:

With a limited number of replications μ cannot be found, only it can be estimated by it's sample mean; $\bar{x}$.

How good is the sample mean in comparison to the (elusive) population mean, μ?

Statistical Evaluation of the mean:

Given the mean $\bar{x}$ and s (or sometimes σ), what is (can be) μ?

Given $\bar{x}$ and s, statistics ⇒ range of values, where μ would be.

Definition: Confidence interval (CI):

Confidence interval is an expression, stating the *range* wherein μ is likely to be found with a certain *degree of confidence*, confidence level (CL),(i.e., with a defined level of reliability).

Confidence level (CL) : *reliability* of the estimate expressed as a percentage (norm - 95%).

Rephrasing 'boxed' question in statisticians' terminology,

Within what values would μ (the population mean) be, so that one can be c% confident that μ is indeed in that interval?

The confidence limits and interval are calculated using,

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}} \quad \text{(confidence limits)}$$

t = Student's t

$$\text{confidence interval} \quad \frac{ts}{\sqrt{n}}$$

$$\text{confidence limits} \quad \left(\bar{x} - \frac{ts}{\sqrt{n}}\right) \leq \mu \leq \left(\bar{x} + \frac{ts}{\sqrt{n}}\right)$$

Confidence limits are the lower and upper limit values where the population mean would be, based on the limited replications.

A 'recipe' to find Confidence Interval/Limits for a data set at c%:

1. Find the value of Student's t from tables relevant to (n-1) degrees of freedom at desired c% confidence.
n = # replications

2. Use t in the formula to find the limits.

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

$$\text{true value} = \mu = \bar{x} \pm \frac{ts}{\sqrt{n}} \quad (n = \ , \text{Conf. Level} = c\%)$$

Larger t, smaller n means wider range of possibilities of the *true value*. Also note that smaller n's are associated with larger t values.

t depends on confidence level & n

Table 4-2 Values of Student's *t*

Degrees of freedom	Confidence level (%) 50	90	95	98	99	99.5	99.9
1	1.000	6.314	12.706	31.821	63.657	127.32	636.619
2	0.816	2.920	4.303	6.965	9.925	14.089	31.598
3	0.765	2.353	3.182	4.541	5.841	7.453	12.924
4	0.741	2.132	2.776	3.747	4.604	5.598	8.610
5	0.727	2.015	2.571	3.365	4.032	4.773	6.869
6	0.718	1.943	2.447	3.143	3.707	4.317	5.959
7	0.711	1.895	2.365	2.998	3.500	4.029	5.408
8	0.706	1.860	2.306	2.896	3.355	3.832	5.041
9	0.703	1.833	2.262	2.821	3.250	3.690	4.781
10	0.700	1.812	2.228	2.764	3.169	3.581	4.587
15	0.691	1.753	2.131	2.602	2.947	3.252	4.073
20	0.687	1.725	2.086	2.528	2.845	3.153	3.850
25	0.684	1.708	2.060	2.485	2.787	3.078	3.725
30	0.683	1.697	2.042	2.457	2.750	3.030	3.646
40	0.681	1.684	2.021	2.423	2.704	2.971	3.551
60	0.679	1.671	2.000	2.390	2.660	2.915	3.460
120	0.677	1.658	1.980	2.358	2.617	2.860	3.373
∞	0.674	1.645	1.960	2.326	2.576	2.807	3.291

n-1

z

t drops with increasing n

NOTE: In calculating confidence intervals, σ may be substituted for *s* in Equation 4-6 if you have a great deal of experience with a particular method and have therefore determined its "true" population standard deviation. If σ is used instead of *s*, the value of *t* to use in Equation 4-6 comes from the bottom row of Table 4-2.

Table 4-2 Values of Student's *t*

Degrees of freedom	Confidence level (%) 50	90	95	98	99	99.5	99.9
1	1.000	6.314	12.706	31.821	63.657	127.32	636.619
2	0.816	2.920	4.303	6.965	9.925	14.089	31.598
3	0.765	2.353	3.182	4.541	5.841	7.453	12.924
4	0.741	2.132	2.776	3.747	4.604	5.598	8.610
5	0.727	2.015	2.571	3.365	4.032	4.773	6.869
6	0.718	1.943	2.447	3.143	3.707	4.317	5.959
7	0.711	1.895	2.365	2.998	3.500	4.029	5.408
8	0.706	1.860	2.306	2.896	3.355	3.832	5.041
9	0.703	1.833	2.262	2.821	3.250	3.690	4.781
10	0.700	1.812	2.228	2.764	3.169	3.581	4.587
15	0.691	1.753	2.131	2.602	2.947	3.252	4.073
20	0.687	1.725	2.086	2.528	2.845	3.153	3.850
25	0.684	1.708	2.060	2.485	2.787	3.078	3.725
30	0.683	1.697	2.042	2.457	2.750	3.030	3.646
40	0.681	1.684	2.021	2.423	2.704	2.971	3.551
60	0.679	1.671	2.000	2.390	2.660	2.915	3.460
120	0.677	1.658	1.980	2.358	2.617	2.860	3.373
∞	0.674	1.645	1.960	2.326	2.576	2.807	3.291

NOTE: In calculating confidence intervals, σ may be substituted for *s* in Equation 4-6 if you have a great deal of experience with a particular method and have therefore determined its "true" population standard deviation. If σ is used instead of *s*, the value of *t* to use in Equation 4-6 comes from the bottom row of Table 4-2.

n-1

n =7
CL = 99%

z-table

CL %	50	90	95	98	99	99.5	99.9
z	0.674	1.645	1.960	2.326	2.576	2.807	3.291

z-table is the last line of the t-table.

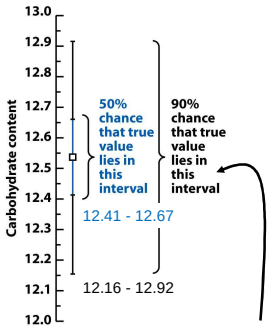
If σ is known from other sources,
then use z-table and *z value* (*which is t* for n = infinity) related to the CL (usually 95%) in the following expression;

$$\mu = \bar{x} \pm z\sigma$$

12.6
11.9
13.0
12.7
12.5

n = 5
n-1 = 4 degree of freedom

Mean =12.5$_4$
s=0.4$_0$



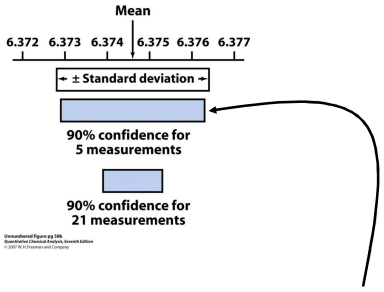
$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$

$$\bar{x} = 12.5$$

$$s = 0.4; \quad n = 5$$

Larger confidence ~ larger t, for same n means a wider range of possibilities for the *true value*. More likely to include μ within the confidence limits.

Speculate the effect of increasing replicates, n, for same CL. (two effects)



For same c%, larger n leads to narrower range of possibilities for the *true value* μ*!! – desired outcome.*

Confidence Interval

Example:
From a Gaussian population of mean μ = 10000 and σ=1000.

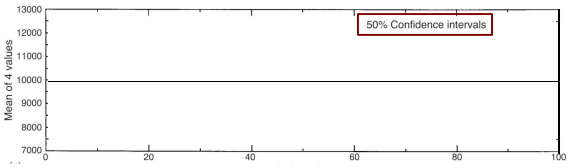
a. Select four data points n = 4 at a time.

b. Calculate the average from the 4 data points its std. deviation and CI at 50% CL.

c. Repeat steps a and b many times.

d. Generate a visual to show the results.

e. Repeat a thro' c for the same data selected but with CL = 90%



Sample data sets (n=4) from the set of population data (points);
population mean= 10000.
■ Confidence limits do not contain μ
☐ Confidence limits contain μ

Standard deviation, standard error (of the mean) and confidence interval estimates are measures of experimental uncertainty.

Reporting result and uncertainty:

Either as Standard Deviation: $result = \bar{x} \pm s$ (n = ..)
note: CL = 68.3%

or

as Standard Error: $result = \bar{x} \pm \frac{s}{\sqrt{n}}$ (n = ..)

or

as a Confidence Limits: $result = \bar{x} \pm \frac{ts}{\sqrt{n}}$ for n = .. and CL = c%

$result = \bar{x} \pm s$ (n = ..)

$result = \bar{x} \pm \frac{s}{\sqrt{n}}$ (n = ..)

$result = \bar{x} \pm \frac{ts}{\sqrt{n}}$ for n = .. and CL = c%

Error bar - objective is to minimize the error bar

Second major use of Student's t:

2. Comparison of Means with Student's t.

Case I; comparison of a measurement with a "known" value

Case IIa; comparing replicates - two data sets - homogeneous - same variances (same protocol)

Case IIb; comparing replicates - two data sets - non-homogeneous – different variances (different protocols)

Case III; comparing individual differences - two data sets - produced by different methods

Student's t:

1. Confidence Interval for a data set at c%:

$result = \bar{x} \pm \frac{ts}{\sqrt{n}}$ for n = .. and CL = c%

2. Comparison of Means with Student's t.

Case I; comparison of a measurement with a "known" value

Case IIa; comparing replicates - two data sets - homogeneous - same variances (same protocol)

Case IIb; comparing replicates - two data sets - non-homogeneous – different variances (different protocols)

Case III; comparing individual differences - two data sets - produced by different methods

Case I (comparison of a measurement with a "known" value)

t-table for validation of a new 'method':

1. Prepare a standard solution of the material, μ.
2. Determine conc. using new method (n replications).
3. Calculate mean, $\bar{x}$ and s for data set.
4. Calculate CI for **95%** confidence level (CL). $\bar{x} \pm \frac{ts}{\sqrt{n}}$
5. If μ falls within CI; ⇒ VALID METHOD.

Case IIa (comparing replicates)

t-tables to compare two data sets
(for same method giving two data sets, implies comparable s values, i.e. homogeneous variances):

Are the data sets significantly different or not?
Strategy: Compare mean values of the data sets.

Requirement: Both sets have the same or nearly the same variances – s^2 (verifiable by F test) i.e., comparable s values.

Set 1: $x_1, x_2, \ldots, x_i$; $\bar{x}_1$ and n_1 observations

Set 2: $x_1, x_2, \ldots, x_j$; $\bar{x}_2$ and n_2 observations

a. Calculate t: t_{calc} for the pooled data from the two data sets.

$$s_{pooled} = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

$$t_{calc} = \frac{|\bar{x}_1 - \bar{x}_2|}{s_{pool}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Calculation of t_{calc} need mean of x_is and n_is.

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

b. Find t-table value for degree of freedom of (n_1+n_2-2) at confidence level of 95% (norm).

c. If $t_{calc} < t_{table}$

Data sets are not significantly different at the confidence level (95%).

t-test: Two samples with equal variances (case IIa) p.64 Harris 7th Edition

Data set 1	Data set 2		t-Test: Two-Sample Assuming Equal Variances		
2.31017	2.30143			*Variable 1*	*Variable 2*
2.30986	2.2989		Mean	2.310108571	2.2994725
2.3101	2.29816		Variance	2.03476E-08	1.90216E-06
2.31001	2.30182		Observations	7	8
2.31024	2.29869		Pooled Variance	1.03363E-06	
2.3101	2.2994		Hypothesized Me	0	
2.31028	2.29849		df	13	
	2.29889		t Stat	20.21372428	Calc t
2.310109	2.299473	Mean	P(T<=t) one-tail	1.66071E-11	
0.000143	0.001379	Std dev	t Critical one-tail	1.770933383	
			P(T<=t) two-tail	3.32141E-11	
			t Critical two-tail	2.160368652	Table t

Case IIb (comparing replicates)

a. If the standard deviations are significantly different, i.e. non-homogeneous, student's t value is calculated using;

$$t_{calc} = |\bar{x}_1 - \bar{x}_2| \sqrt{\frac{1}{(s_1^2/n_1)+(s_2^2/n_2)}}$$

and the degrees of freedom (to the nearest integer) with;

$$DF = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1}\right)} - 2$$

b. Find the t-table value for degree of freedom (DF) calculated, at specified confidence level of 95% (norm)

c. If $t_{calc} < t_{table}$

Data sets are not significantly different at the confidence level (95%)

t-test: Two samples with unequal variances (case IIb) p.64 Harris 7th Edition

Data set 1	Data set 2		t-Test: Two-Sample Assuming Unequal Variances			
2.31017	2.30143			*Variable 1*	*Variable 2*	
2.30986	2.2989		Mean	2.310108571	2.2994725	
2.3101	2.29816		Variance	2.03476E-08	1.90216E-06	
2.31001	2.30182		Observations	7	8	
2.31024	2.29869		Hypothesized Me	0		
2.3101	2.2994		df	7		Calculated degrees of freedom
2.31028	2.29849		t Stat	21.68021802		Calculated Students t
	2.29889		P(T<=t) one-tail	5.60174E-08		
2.310109	2.299473	Mean	t Critical one-tail	1.894578604		
0.000143	0.001379	Std dev	P(T<=t) two-tail	1.12035E-07		
			t Critical two-tail	2.364624251		Table value of Students t

Case III (comparing individual differences)

t-tables to compare two methods i.e. two data sets produced by the two different methods:

a. Subject two sets of laboratory samples (n replicates each) to the two protocols (methods).

b. Tabulate the results for each sample

Sample label	Method 1 (unit)	Method 2 (unit)	Difference d_i
1	17.2	14.2	-3
2	23.1	27.9	4.8
3	28.5	21.2	-7.3
4	15.3	15.9	0.6
5	23.1	32.1	9
6	32.5	22	-10.5
7	39.5	37	-2.5
8	38.7	41.5	2.8
9	52.5	42.6	-9.9
10	42.6	42.8	0.2
11	52.7	41.1	-11.6

$t_{calc} = \frac{|\bar{d}|}{s_d}\sqrt{n}$

$s_d = \sqrt{\frac{\sum_i^n (d_i - \bar{d})^2}{n-1}}$

-2.49091 = mean

6.748252 = std dev

1.22423 = t_calc

t_{table}= 2.228 two methods produce same results

c. For the differences

$$s_d = \sqrt{\frac{\sum_i^n (d_i - \bar{d})^2}{n-1}} \qquad t_{calc} = \frac{|\bar{d}|}{s_d}\sqrt{n}$$

d. Find t_{table} for (n-1) deg. of freedom at 95% CL.

e. If $t_{calc} < t_{95\%,table}$; the methods produce results which are not significantly different at 95% CL.

Table 4-2 Values of Student's *t*

Degrees of freedom	Confidence level (%) 50	90	95	98	99	99.5	99.9
1	1.000	6.314	12.706	31.821	63.657	127.32	636.619
2	0.816	2.920	4.303	6.965	9.925	14.089	31.598
3	0.765	2.353	3.182	4.541	5.841	7.453	12.924
4	0.741	2.132	2.776	3.747	4.604	5.598	8.610
5	0.727	2.015	2.571	3.365	4.032	4.773	6.869
6	0.718	1.943	2.447	3.143	3.707	4.317	5.959
7	0.711	1.895	2.365	2.998	3.500	4.029	5.408
8	0.706	1.860	2.306	2.896	3.355	3.832	5.041
9	0.703	1.833	2.262	2.821	3.250	3.690	4.781
10	0.700	1.812	2.228	2.764	3.169	3.581	4.587
13	0.691	1.753	2.131	2.602	2.947	3.252	4.073
20	0.687	1.725	2.086	2.528	2.845	3.153	3.850
25	0.684	1.708	2.060	2.485	2.787	3.078	3.725
30	0.683	1.697	2.042	2.457	2.750	3.030	3.646
40	0.681	1.684	2.021	2.423	2.704	2.971	3.551
60	0.679	1.671	2.000	2.390	2.660	2.915	3.460
120	0.677	1.658	1.980	2.358	2.617	2.860	3.373
∞	0.674	1.645	1.960	2.326	2.576	2.807	3.291

n-1

NOTE: In calculating confidence intervals, σ may be substituted for *s* in Equation 4-6 if you have a great deal of experience with a particular method and have therefore determined its "true" population standard deviation. If σ is used instead of *s*, the value of *t* to use in Equation 4-6 comes from the bottom row of Table 4-2.

Comparison of two methods

Sample label	Method 1 (unit)	Method 2 (unit)	Difference d_i
1	17.2	14.2	-3
2	23.1	27.9	4.8
3	28.5	21.2	-7.3
4	15.3	15.9	0.6
5	23.1	32.1	9
6	32.5	22	-10.5
7	39.5	37	-2.5
8	38.7	41.5	2.8
9	52.5	42.6	-9.9
10	42.6	42.8	0.2
11	52.7	41.1	-11.6

t-Test: Paired Two Sample for Means

	Variable 1	Variable 2
Mean	33.24545	30.75455
Variance	171.4027	121.9507
Observatio	11	11
Pearson C	0.857029	
Hypothesiz	0	
df	10	
t Stat	1.22423	
P(T<=t) on	0.124462	
t Critical o	1.812461	
P(T<=t) tw	0.248924	
t Critical tv	2.228139	

The F test

Some statistical procedures calls for pooling of variances, (Case IIa & IIb). Here a knowledge of the nature of the sample populations is needed i.e. are variances homogeneous or not. (similar s values or not).

So before proceeding with a procedure that pools variances, it is necessary to test for the assumption of homogeneity of variances

The **F-test** provides a tool for comparing variances of data sets.

In such a test, the outcome should be either there is no difference in population variances (Null hypothesis) or there is a difference in the population variances (Alternative hypothesis).

The test is performed using the variances of the two data sets, say, set 1 (#1 assigned to the set with *larger* s) and set 2.

$n_2 = 8$ $\quad n_1 = 7$

$s_2 = 0.001379$ $\quad s_1 = 0.001430$

Are the variances (s^2) and therefore s values different?

One of the variances of the two data sets should be larger than the other. Let the set 1 to be the one with larger variance s_1^2 (> s_2^2). Calculate the fraction $F_{calc} = F_{expt}$, written as follows;

$$F_{calc} = F_{(n_1, n_2)} = \frac{s_1^2}{s_2^2}$$

Now compare the F_{calc} to the F_{crit} from a critical F values (table), where the degrees of freedom are (n_1 - 1), (n_2 - 1), (n_1 and n_2 are the number of replicates in data sets 1 and 2).

Reject the null hypothesis if the $F_{calc} > F_{crit}$. That is the variances are different.

There is no difference in the variances, if $F_{calc} < F_{crit}$

Symbols used $F_{crit} = F_{table}$; $F_{expt} = F_{calc}$

$n_2 = 8$ $\quad$ $n_1 = 7$

$s_2 = 0.001379$ $\quad$ $s_1 = 0.001430$

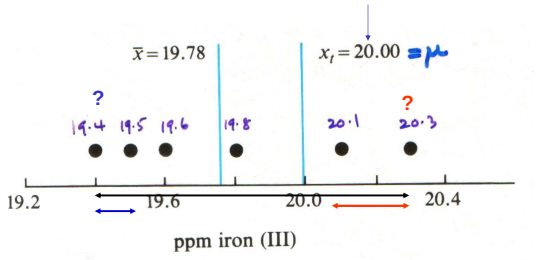
$$F_{calc} = \frac{s_1^2}{s_2^2} = \frac{0.001430^2}{0.001379^2} = 1.075$$

$F_{table} = 3.87 > F_{calc}$ variances not different!

Subscript 1 associated with larger s.

Table 4-5 Critical values of $F = s_1^2/s_2^2$ at 95% confidence level

Degrees of freedom for s_2	Degrees of freedom for s_1													
	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
2	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5
3	9.55	9.28	9.12	9.01	8.94	8.89	8.84	8.81	8.79	8.74	8.70	8.66	8.62	8.53
4	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.63
5	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.36
6	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.67
7	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.58	3.51	3.44	3.38	3.23
8	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	2.93
9	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.71
10	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.70	2.54
11	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.40
12	3.88	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.30
13	3.81	3.41	3.18	3.02	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.21
14	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.13
15	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.07
16	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.01
17	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	1.96
18	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	1.92
19	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	1.88
20	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.84
30	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.62
∞	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.00



Outliers, if exists appear at the extremes.

Rejection of outliers: Q Test

Outliers are not always obvious. To reject a suspicious data point from set of n data points, where there is no obvious gross error, the Q test is used.

a. Arrange the data in the order of increasing value.

b. Determine the range = (x_{max} - x_{min})

c. Find the difference between the data point in question, and its nearest neighbor. gap = $|X_q - X_n|$

d. Calculate the rejection quotient Q_{calc} as; $Q_{calc} = \frac{\text{gap}}{\text{range}}$

e If $Q_{calc} < Q_{table}$ for the n, accept x_q for a given confidence level, 90% - norm. (> i.e 10% chance it is an outlier).

Table 4-6 Values of Q for rejection of data

Q (90% confidence)[a]	Number of observations
0.76	4
0.64	5
0.56	6
0.51	7
0.47	8
0.44	9
0.41	10

a. Q = gap/range. If $Q_{calculated} > Q_{table}$, the value in question can be rejected with 90% confidence.

SOURCE: R. B. Dean and W. J. Dixon, *Anal. Chem.* **1951**, *23*, 636; see also D. R. Rorabacher, *Anal. Chem.* **1991**, *63*, 139.

Rejection of outliers: Grubbs Test

Calculate the Grubbs statistic.

$$G_{calc} = \frac{\left|questionable\ value - \bar{x}\right|}{s}$$

Compare G_{calc} vs Critical table values for G for n observations

If $G_{calc} < G_{table}$; accept the questionable value at 95% CL.

TABLE 4-5 Critical values of G for rejection of outlier

Number of observations	G (95% confidence)
4	1.463
5	1.672
6	1.822
7	1.938
8	2.032
9	2.110
10	2.176
11	2.234
12	2.285
15	2.409
20	2.557

$G_{calculated}$ = |questionable value − mean|/s. If $G_{calculated} > G_{table}$, the value in question can be rejected with 95% confidence. Values in this table are for a one-tailed test, as recommended by ASTM.

SOURCE: ASTM E 178-02 Standard Practice for Dealing with Outlying Observations, http://webstore.ansi.org; F. E. Grubbs and G. Beck, Technometrics **1972,** *14,* 847.

Harris, *Quantitative Chemical Analysis,* 8e

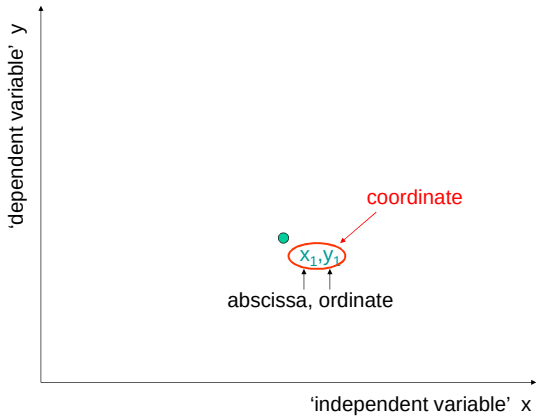
Graphs:

Graphs are an essential and efficient way to communicate experimental data in a visual manner.

Normally, relationships between two variables are plotted; 'independent variable' on x axis and dependent variable on y axis.

Whenever possible, linearization of functions is done. *Linear functions allow better mathematical algorithms to find the best fit line*. Linearization is not essential if a well established mathematical relationship (non-linear) is available relating the [x,y] coordinates.

In analytical chemistry – (external) calibration plots, internal standard (addition) plots, and standard addition plots; (straight lines) are widely used in quantification.



Least Square Analysis - Linear Regression Plots:

Often the desired quantities are determined from line graphs.

General form; y = m x + b

Regression plots are used as calibration 'curves' as well.

Calibration curve: A graph showing the variation of the value of a property, y, as a function of known analyte concentration, c, under specified conditions.

E.g. y = m c + b

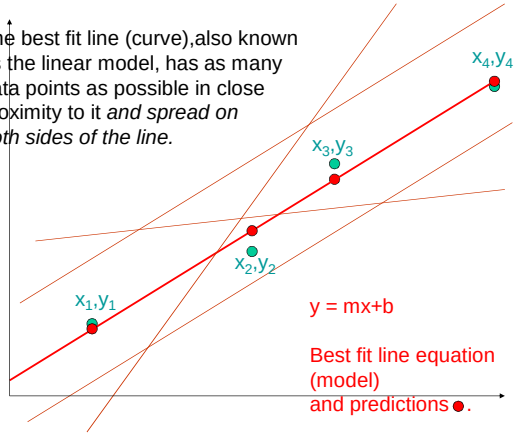
The plot is usually constructed from the y-values corresponding to a set of standard solutions (concentration, x-values, known) - external calibration curve.

The y-value, y_u (mean of k replicates) associated to an unknown conc. x_u, is then used to determine unknown conc. x_u and it's uncertainty.

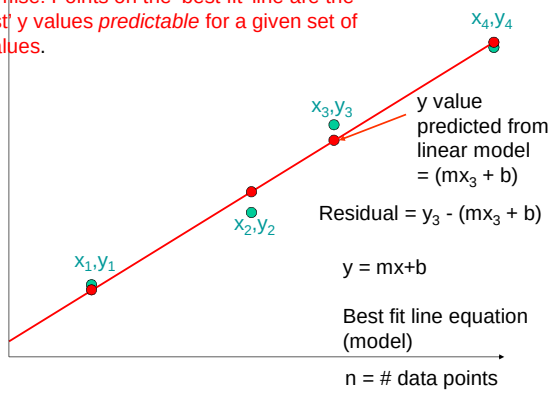
Experiment with no uncertainties produces data points, all on a perfect function, say a straight line. But, in real experimentation such outcomes are very very rare.

Usually, the data points are not exactly on a line, but scattered around a line.

The best fit line (curve),also known as the linear model, has as many data points as possible in close proximity to it *and spread on both sides of the line.*



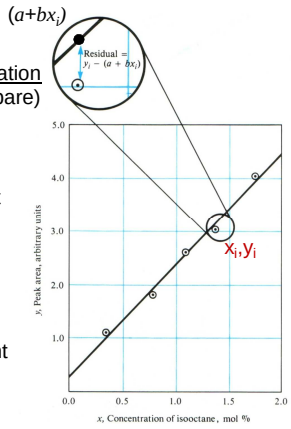
Premise: Points on the 'best fit' line are the 'best' y values *predictable* for a given set of x values.

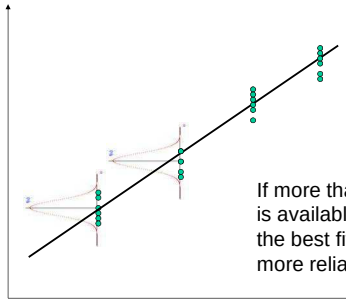


A <u>single point per concentration</u> in a calibration plot is the (bare) minimum number of data points possible.

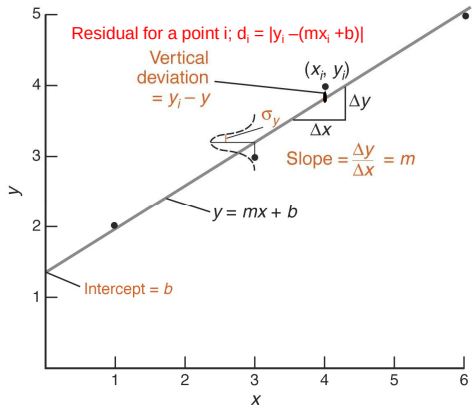
If more than one data point is obtained the resulting data points would have a spread of values.

Every single data point is very u*nlikely* to be the 'right on' the line.





If more than one data point is available for each x value the best fit line would be more reliable.



Data points deviates from the straight line, smaller the deviation better is the precision.

How to determine line with least deviations? i.e. best fit line.

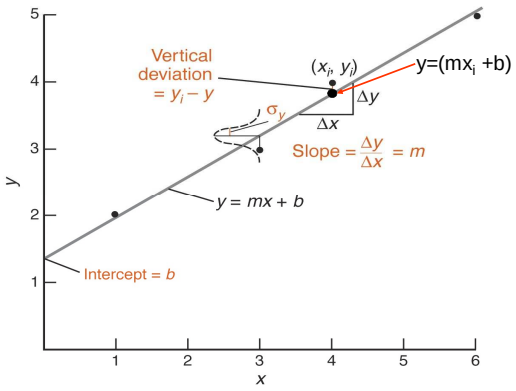
<u>LSA:</u> Least Square Analysis.

Objective (of LSA) – to find the best straight line relevant to the data set – best line fit.

LSA strategy minimizes the residuals for all points.
For data point i; $d_i = |y_i - (mx_i + b)|$

Assumption: x values are 'exact'
y values contain an error.

Best line – produces values for m and b that would minimize all deviations/residuals (positive or negative).



Criterion: find the line so that the sum of squares of residuals for the data set is a minimum (least square minimization).

Mathematical treatment of a data set will generate the 'best vales' for m and b; 'the best fit line parameters' (Eq 4-16, 4-17).

LSA Assumptions:

- Uncertainties in x (e.g. standard concentrations) negligible compared to those of y (observations) values.
- Uncertainties in y (observations; e.g. absorbance) values are similar.

The uncertainty/errors (std. deviation) for slope, intercept and x_u calculated from the calibration plot.

The best fit lines are generated with the data points (x_i, y_i) of a data set and each point is associated with an uncertainty.

So the parameters in the best fit line and the quantities calculated using the best fit line are associated with uncertainties (standard deviations, s_i).

slope, m	s_m
intercept, b	s_b
calculated , x_u	s_u
Overall y	s_y

FYI

$$\sigma_y \approx s_y = \sqrt{\frac{\sum d_i^2}{n-2}} \qquad d_i = | y_i - (mx_i + b) |$$

n = #data points (calibration)

k = # replicates (unknown)

$$s_m^2 = \frac{s_y^2 n}{D}$$

$$s_b^2 = \frac{s_y^2 \sum_i x_i^2}{D}$$

$$D = n\sum_i x_i^2 - \left(\sum_i x_i\right)^2$$

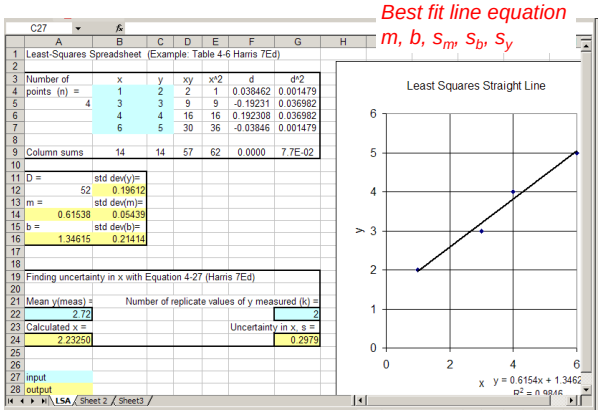
$$s_u = \frac{s_y}{|m|}\sqrt{\frac{1}{k} + \frac{1}{n} + \frac{(y_u - \bar{y})^2}{m^2 \sum (x_i - \bar{x})^2}}$$

$$S_{xx} = s_x^2 (n-1)$$

Mathematically accurate expressions for m, b and s_m, s_b and s_y.

Least Squares Spreadsheet

y = mx+b
Best fit line equation
m, b, s_m, s_b, s_y



Concentration of unknown, u, is calculated using the calibration curve (at hand).

a. find the y_u values for k replications of the unknown. Calculate the mean of y_u.

b. calculate the concentration, mean x_u values corresponding to y_u values in (a).

The error for x_u (usual notation x) from a best fit line calibration curve, with s_b, s_m values defined below and *k* replications of the unknown be s_x.

n = # of calibration points used for the calibration curve.

Result = $x_u \pm s_u$.

Confidence interval of x_u; $x_u \pm (t_{\%p} s_u)/\sqrt{k}$ at p% and k = : better error bar

$t_{\%p}$ for (n-2) degrees of freedom (if using best fit line) at p% confidence level (%p = 95% usually where n = # of calibration points and k = number of replicated measurements).

See p.72 for an EXCEL routine to calculate s_m, s_b, s_x and s_y.

The overall objective is to obtain *an accurate estimate of an unknown with the smallest possible uncertainty*.
To minimize the uncertainty;

Make s_y small – a good fit
Obtain many k replicate measurements
A sufficiently large range for calibration *and* accurate standards.
Signal of the 'unknown' in the 'middle of the calibration range'.
Increase the number of calibration data points, n.

Optimum for k = n and n > 5 (⇒ smaller t-values)

$$s_x = s_u = \frac{s_y}{|m|}\sqrt{\frac{1}{k}+\frac{1}{n}+\frac{(y_u-\bar{y})^2}{m^2\sum(x_i-\bar{x})^2}}$$

Table 5-2 Spectrophotometer data used to construct calibration curve

Amount of protein (μg)	Absorbance of independent samples			Range	Corrected absorbance		
0	0.099	0.099	0.100	0.001	-0.000_3	-0.000_3	0.000_7
5.0	0.185	0.187	0.188	0.003	0.085_7	0.087_7	0.088_7
10.0	0.282	0.272	0.272	0.010	0.182_7	0.172_7	0.172_7
15.0	0.345	0.347	~~(0.392)~~	0.047	0.245_7	0.247_7	—
20.0	0.425	0.425	0.430	0.005	0.325_7	0.325_7	0.330_7
25.0	0.483	0.488	0.496	0.013	0.383_7	0.388_7	0.396_7

data points on calibration curve, *n* = 14

Correlation coefficient, R:

The coefficient assesses the degree of linearity between two variables y and x.

$$R = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\left[n\sum x_i^2 - (\sum x_i)^2\right]^{1/2}\left[n\sum y_i^2 - (\sum y_i)^2\right]^{1/2}}$$

$$= \frac{\sum_1^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_1^n (x_i-\bar{x})^2 \sum_1^n (y_i-\bar{y})^2}}$$

R = +1; positive slope ideal fit ; $R^2 = 1$
R= -1; negative slope ideal fit; $R^2 = 1$
R = 0; zero slope ideal fit; $R^2 = 0$

Least squares method considers all data points to calculate the m and b of the line.

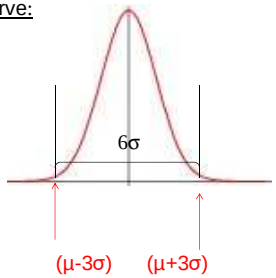
Therefore, outliers skew the 'best fit line'.

Each outlier, which by definition, is a point $3\sigma_y$ or $3s_y$ away from the 'line' must be identified and eliminated; and then the line fitting is redone with the remaining data points.

This process must be done iteratively to identify the outliers.

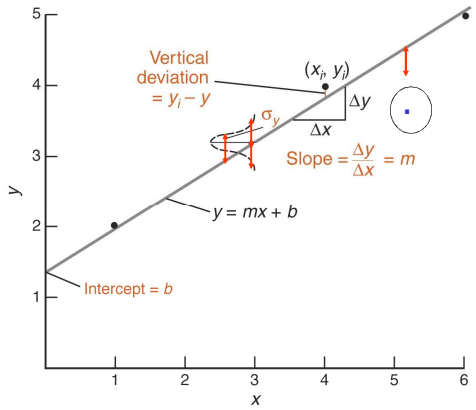
An alternative method is the Thiel-Siegel line fitting method.
Leslie Glasser**,** J Chem Ed, Vol. 84 , 533, 2007

Gaussian Curve:

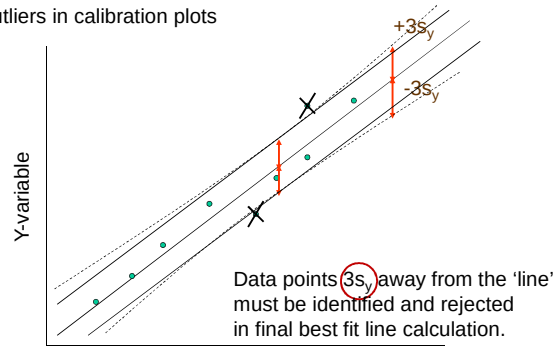


The area within the 6σ limit is 99.7% of the total area.

$\text{mean} \pm \sigma$	68.3%;
$\text{mean} \pm 2\sigma$	95.4%
$\text{mean} \pm 3\sigma$	99.7%

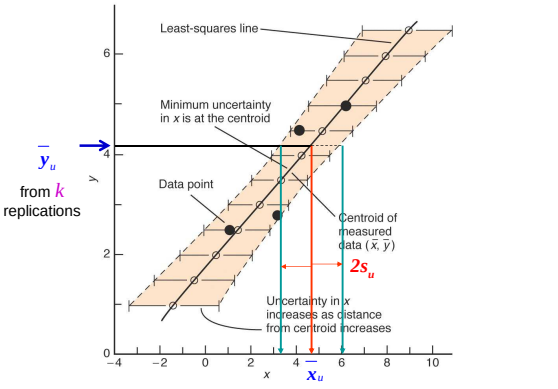


Outliers in calibration plots

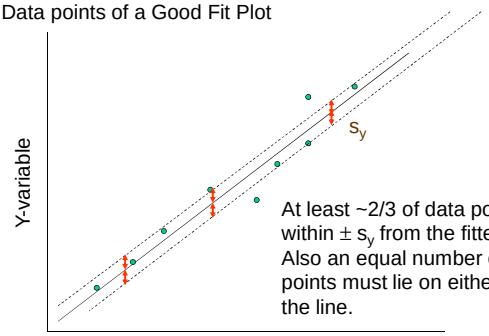


Data points 3s_y away from the 'line' must be identified and rejected in final best fit line calculation.

$$s_x = s_u = \frac{s_y}{|m|}\sqrt{\frac{1}{k} + \frac{1}{n} + \frac{(y_u - \bar{y})^2}{m^2 \sum (x_i - \bar{x})^2}}$$



Data points of a Good Fit Plot



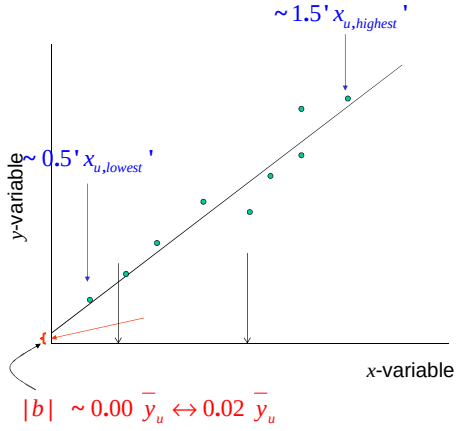
At least ~2/3 of data points lies within ± s_y from the fitted line. Also an equal number of data points must lie on either side of the line.

Calibration curve *range* preferably includes the 'test' concentrations; preferably *0.5 to 1.5 times* the 'test' concentrations.

Each standard – run in at least in triplicate.

$R^2 > 0.995$ good fit.

y-intercept, |b| – (after correcting for the blank) < 2% y target value of 'test'.



Thiel-Siegel Line Fitting

Data 1 from Massart, Table 8.7 (with duplicate and empty cells)			Least-squares (LS)
Theil/Siegel	Slope, b	1.017E+00	1.691
	Intcpt, a	2.500E-02	-0.895
	No. data	6	
X	Y	T/S fitted Y	(LS resid.)^2
0	0.000	0.025	0.80
1	1.100	1.042	0.09
2	2.000	2.058	0.24
3	3.100	3.075	1.16
4	3.800	4.092	4.29
5	10.000	5.108	5.94