

Chemometrics

Unit 1. Measurement and Assessment of Variability

1.1 Description of Variability

One of the most common tasks an analytical chemist carries out is to make repeated measurements of a quantity. From these measurements estimates can be made of the true value of the quantity being measured and the reliability of this estimate. The criteria used to evaluate the data are:-

$$\text{Arithmetic Mean, } \bar{x}, = \frac{\sum x_i}{n}$$

$$\text{and Variance, } s^2, = \frac{\sum (x_i - \bar{x})^2}{(n-1)}$$

s is also called the **standard deviation**.

Another quantity commonly used is the **coefficient of variation (CV)** or **Relative Standard Deviation (RSD)**,

$$\text{RSD} = \frac{100s}{\bar{x}}$$

This gives an idea of the error of the determination expressed as a percentage.

1.2 The Normal Distribution.

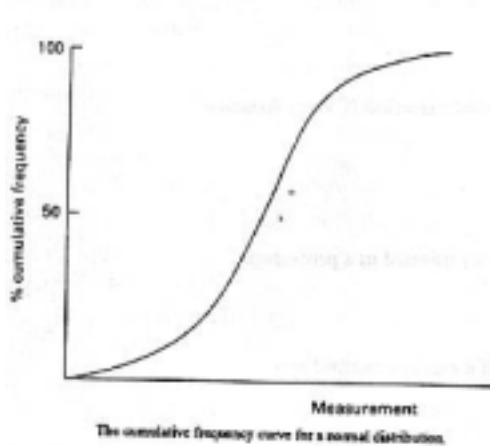
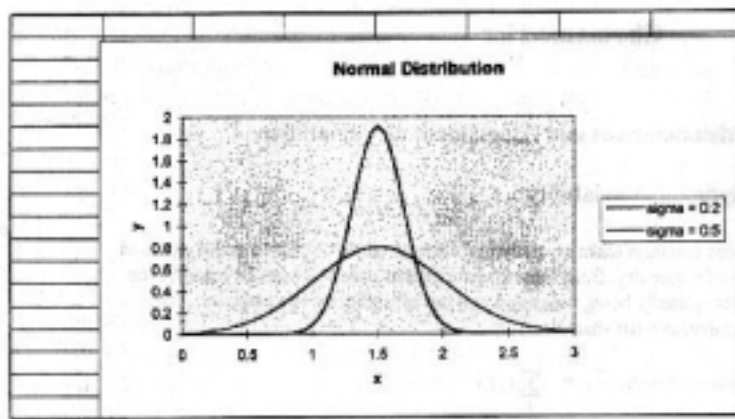
A variable, y , is said to be **normally distributed** if it can be described by:-

$$y = \frac{\exp(-(x-\mu)^2 / 2\sigma^2)}{\sigma\sqrt{2\pi}}$$

The shape of the curve (fig 1) is completely determined by the two parameters μ and σ . The maximum of the curve is at $x = \mu$ and the 'fatness' of the curve is determined by σ . (Fig 1 shows curves for $\sigma = 0.2$ and $\sigma = 0.5$).

1.2.1 Testing for Normality of a Distribution.

Many statistical tests assume that data used has been drawn from a normally distributed population. It is important to be able to test whether data does fit a normal distribution. One way of testing whether a set of data is consistent with the assumption of normality is to plot a **cumulative frequency curve** on normal probability paper.



Scale for producing normal plots of effects and residuals

Fig. 2a

Fig. 2b

Example: A student carries out a series of titrations and gets these results:

25.02, 24.90, 25.12, 25.05, 24.98, 25.26, 25.10, 24.96.

Are the titrations normally distributed can any points be regarded as outliers?

The first step is to arrange the data in order of increasing magnitude. The cumulative frequency is then calculated:

Titre	Cumulative Frequency	% Cumulative Frequency
24.90	1	11.1
24.96	2	22.2
24.98	3	33.3
25.02	4	44.4
25.05	5	55.6
25.1	6	66.7
25.12	7	77.8
25.26	8	88.9

where % cumulative frequency = $100 * (\text{cumulative frequency}) / (n+1)$; n is the number of measurements.

If this data is plotted (titre vs cumulative frequency) and s shaped curve is expected if the data is normally distributed (fig 2a). Normal probability paper has a non-linear scale for the frequency axis (fig 2b) which converts the S-shaped curve into a straight line. Fig 3 shows a plot of the titration data.

Normal probability plots can be produced using MINITAB (choose probability plot from the *Graph Menu*). The plot for the titration data is shown in fig 3. It can be seen that the data is linear, except for the 25.26 value. Thus this point can be considered as an outlier. We would accept the hypothesis that the rest of the data is normally distributed.

Normal distributions are particularly important when examining residuals in regression analysis (see Unit 2).

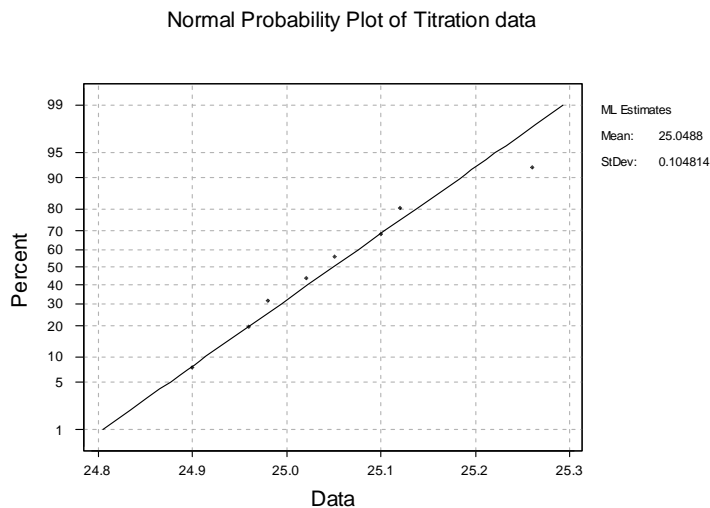


Fig. 3

1.3 The Null Hypothesis

In any statistical test we first need a hypothesis to test and then a way of evaluating the truth or otherwise of this hypothesis. Central to this is the concept of the **null hypothesis** (abbreviated H_0). Generally, H_0 is the hypothesis that there is **no** systematic differences between the data i.e. differences can be accounted for by random chance.

The alternative hypothesis, H_1 , can take several forms and the way the test is evaluated depends on the form of H_1 .

Type 1 error: reject H_0 even if it is true

Type 2 error: accept H_0 although it is false

Generally the tendency is to err on the side of caution so type 2 errors may be preferred to type 1 (although in toxicology measurements type 1 errors may be preferable).

Example: A test is carried out to see if chemical X may be carcinogenic, by feeding it to test animals.

H_0 : the incidence of cancer in the test animals is not significantly different to the controls (i.e. X is not carcinogenic).

A type 1 error would be to reject H_0 i.e. to accept that X is carcinogenic when it really isn't. A type 2 error would be to claim X is safe although it is carcinogenic.

1.4 Tests of Significance: The 'Student's t' Test

1.4.1 Comparison of Experimental Mean with Known Value.

An important example of a normally distributed variable is the t statistic. We can use this statistic to evaluate the null hypothesis: $H_0: x = \mu$ by calculating t as follows:

$$t = \frac{(\bar{x} - \mu)}{s / \sqrt{n}}$$

Alternative hypothesis: $H_1: x \neq \mu$

H_0 is rejected if $|t|$ exceeds a critical value. The critical value depends on the **confidence level**, α , (defined as the % chance of H_0 being true - generally if there is a greater than 5% chance of H_0 being true, H_1 is rejected) and the **degrees of freedom**, ν (the number of **independent measurements**).

H_1 here is an example of a **two-tailed** test. A **one-tailed** test would be: $H_1: x > \mu$. The critical value of t depends on the form of H_1 .

Example: An analyst carries out an analysis of mercury in a piece of fish. Replicate determinations give the following results:

0.52, 0.57, 0.51, 0.54, 0.55. The legal limit is 0.5. Can we regard this sample as **significantly** above the limit?

$$\bar{x} = (0.52 + 0.57 + 0.51 + 0.54 + 0.55)/5 = 0.538$$

$$s^2 = 0.25 * \{(.52-.538)^2 + (.57-.538)^2 + (.51-.538)^2 + (.54-.538)^2 + (.55-.538)^2\}$$

$$= 0.00057; \quad s = 0.0239$$

$$|t| = |(.538 - .5)| / (.0239 / 5^{0.5}) = 3.56$$

This is a one-tailed test so at the 5% level of significance ($\alpha = 0.05$) and, using table 1, we need to consult the $\alpha = 0.1$ column. The number of degrees of freedom ($n - 1$) is 4 and $t_{0.1,4} = 2.13$. 3.56 is greater than 2.13 so we **reject** H_0 i.e. the amount of mercury **is** significantly greater than 0.5.

The analysis can be carried out in MINITAB . Choose Basic Statistics > 1-sample t. Check the 'expected mean' box and enter '0.5'

T-Test of the Mean

Test of $\mu = 0.5000$ vs $\mu > 0.5000$

Variable	N	Mean	StDev	SE Mean	T	P
C1	5	0.5380	0.0239	0.0107	3.56	0.012

- (i) This is a one-tailed test so 'greater than' is specified..
- (ii) The default for MINITAB is 95% level of confidence (i.e $\alpha = 0.05$)
- (iii) MINITAB also includes a **probability**, p , of H_0 being true. Since $p < 0.05$ we **reject** H_0 .

1.4.2 Confidence Intervals for the Mean

The t statistic can be used to evaluate confidence intervals for the mean:-

$(\bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}})$ where $t_{\alpha, v}$ is obtained from tables ($\alpha = 0.05, v = n-1$). This interval gives a 95% chance that the 'real' mean, μ , lies in this interval.

In the above Hg example, an alternative way to test H_0 would be to see if 0.5 lies in the 95% C.I. for the data (enter tinterval C1 in the command line editor):

(MINITAB output)

T Confidence Intervals

Variable	N	Mean	StDev	SE Mean	95.0 % CI
C1	5	0.5380	0.0239	0.0107	(0.5084, 0.5676)

The C.I. doesn't include 0.5 so reject H_0 .

1.4.3 Comparison of Two Means

The t-statistic can be used to test whether two means differ significantly:-

$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s^* \sqrt{1/n_1 + 1/n_2}}$; t has $n_1 + n_2$ degrees of freedom, where n_1 and n_2 are the numbers of samples in each set. s^* is a pooled standard deviation, where

$$s^{*2} = \frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1 + n_2 - 2}$$

Example (M & M p 56) In a series of experiments to determine tin in food, samples were refluxed in HCl for different times and these results were obtained:-

Reflux Time (min)	[Sn] mg/kg
30	55, 57, 59, 56, 59
75	57, 55, 58, 59, 59

$$x_1 = 57.0; \quad x_2 = 57.53; \quad s_1^2 = 2.8; \quad s_2^2 = 2.57$$

$H_0: x_1 = x_2$ i.e reflux time has no effect on results. $H_1: x_1 \neq x_2$

$$s^{*2} = (5*2.8 + 2*2.7)/10 = 2.685; \quad s = 1.64$$

$$t = \frac{(57.0-57.83)}{1.64\sqrt{1/6+1/6}} = -0.88; \quad t_{.05,10} = 2.23$$

$|t| = .88 < 2.23$ - cannot reject H_0 , so there is no evidence that reflux time affects the result. The test is carried out in Minitab by Basic Statistics > 2-sample t

Two Sample T-Test and Confidence Interval

Two sample T for R30 vs R75

	N	Mean	StDev	SE Mean
R30	5	57.20	1.79	0.80
R75	5	57.60	1.67	0.75

95% CI for μ R30 - μ R75: (-2.99, 2.19)

T-Test μ R30 = μ R75 (vs not =): T = -0.37 P = 0.73 DF = 7

$p = 0.4$ of H_0 being true, > 0.05 accept H_0

1.5 Comparison of Variance: the F-Test

The F-test considers the ratio of variance: $F = \frac{S_1^2}{S_2^2}$ (s_1 and s_2 allocated so $s_1 > s_2$ i.e F

> 1). The F test can be used to compare the **precision** of two sets of data.

Example: in the above example of the tin analyses, we can test whether the reflux times affect the **precision** of the analyses:

$F = 2.8/2.57 = 1.09$ For $H_0: s_1^2 = s_2^2$ Checking (table 2) for $\alpha = 0.05$, $v_1 = n_1 - 1 = 5$, $v_2 = n_2 - 1 = 5$ $F_{.05,5,5} = 7.15$. Since 1.09 is less than 7.15 we can accept H_0 i.e. no differences in precision.

1.6 Comparison of Several Means: Analysis of Variance (ANOVA)

The t-test enables comparison of two means but what if we want to compare three sets of data? 3 t tests would be needed (A vs B, A vs C and B vs C) and the number of comparisons rapidly increases as the number of sets increases (6 for 4 sets, 10 for 5 sets, 15 for 6 sets etc). Analysis of Variance gives a more efficient way of evaluating differences between sets of data.

Example (M&M p 66) The data shown is from an investigation into the stability of a fluorescent reagent when stored :-

H_0 : no change on storage

Conditions	Replicates	Mean
A Freshly Prepared	102,101,101	101
B Stored for 1 hr in dark	101,101,014	102
C Stored for 1 hr in subdued light	97,95,99	97
D Stored for 1 hr in bright light	90,92,94	92

Steps: 1. Evaluate **total** variability (total sum-of-squares)

$$\sum_{i=1}^r \sum_{j=1}^c (y_{ij} - y)^2 \quad r = \text{treatments (4); } c = \text{replicates (3) } r*c = N = 12$$

$$\text{average } y = y^* = 98$$

$$= (102 - 98)^2 + (100 - 98)^2 + (101 - 98)^2 + (101 - 98)^2 + (101 - 98)^2 + (104 - 98)^2 + (97 - 98)^2 + (95 - 98)^2 + (99 - 98)^2 + (90 - 98)^2 + (92 - 98)^2 + (94 - 98)^2$$

$$= 210 = \text{SST}$$

2. Split the variability into contributions due to differences between samples (treatments) and within samples

(a) Between samples

$$\text{SSA} = c * \sum_{i=1}^r (y_i - y)^2$$

where y_i is the **average** for treatment i

$$= 3 * \{ (101 - 98)^2 + (102 - 98)^2 + (97 - 98)^2 + (92 - 98)^2 \}$$

$$= 3 * 62 = 186$$

(b) **Within Samples:** this is an estimate of the **error** of the method:

$$\text{SSE} = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - y_i)^2 \quad \text{where } y_i \text{ is the means of each treatment}$$

$$= (102 - 101)^2 + (100 - 101)^2 + (101 - 101)^2 + (101 - 102)^2 + (101 - 102)^2 + (104 - 102)^2 + (97 - 97)^2 + (95 - 97)^2 + (99 - 97)^2 + (90 - 92)^2 + (92 - 92)^2 + (94 - 92)^2$$

$$= 24$$

Step 2: The sum-of-squares are divided by the degrees of freedom

$$\text{SSA/DofF} = 180 / (4 - 1) = 62$$

$$\text{SSE/DofF} = 24 / (4 * (3 - 1)) = 3$$

Step 3: The F test is used to compare the between treatments variability to the error

$$F = 62 / 3 = 20.67 \quad F_{0.05, 3, 8} = 5.42$$

$20.67 > 5.42$ so we reject H_0 .

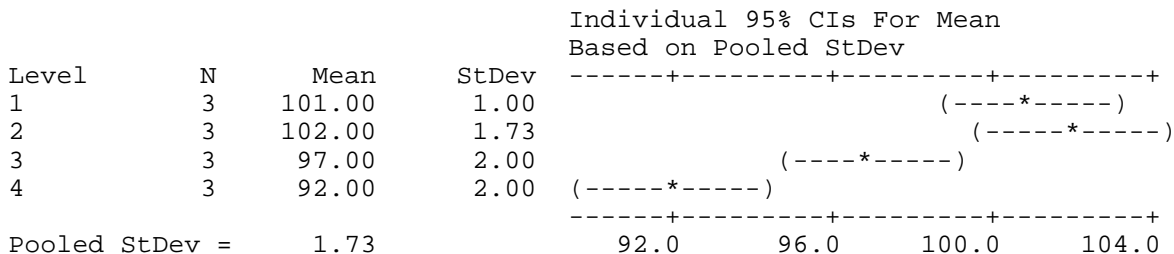
(Minitab) There are two commands that can be used - **oneway (unstacked)** and **oneway**. They only differ in the way the data is entered. In **oneway (unstacked)** each treatment is in a separate column. In **oneway** all the data is in one column and a second column contains a code to label which treatment (can be any numbers but usually 1 = A or treatment 1, 2 = B or treatment 2 etc). Columns C1 and C2 show the structure for **oneway**, C3-C6 for **oneway (unstacked)**

	C1	C2	C3	C4	C5	C6
	Fsignal	storage	A	B	C	D
1	102	1	102	101	97	90
2	100	1	100	101	95	92
3	101	1	101	104	99	94
4	101	2				
5	101	2				
6	104	2				
7	97	3				
8	95	3				
9	99	3				
10	90	4				
11	92	4				
12	94	4				

One-way Analysis of Variance

Analysis of Variance for Fsignal

Source	DF	SS	MS	F	P
Storage	3	186.00	62.00	20.67	0.000
Error	8	24.00	3.00		
Total	11	210.00			



Analysis of Variation for One Way Classification

Source of	Sum of Squares	Degrees of	Mean Square	Computed
-----------	----------------	------------	-------------	----------

Variation		Freedom		
Treatments	SSA	k-1	$s_1^2 = SSA/(k-1)$	$F = s_1^2/s^2$
Error	SSE	k(n-1)	$s^2 = SSE/(k(n-1))$	
Total	SST	nk-1		

1.7 Two-Way Analysis of Variance

Consider now the case where we have more than one source of variation.

Example: An experiment to compare the % efficiency of different chelating agents in extracting metal ions from aqueous solutions gave the following results: (each extraction was performed in duplicate)

	Chelating Agent			
Day	A	B	C	D
1	84,85	79,81	83,83	77,79
2	79,79	76,78	79,81	79,80
3	83,86	78,79	80,79	77,78

On each day a fresh solution of metal ion was prepared. This is known as **blocking** the experiment. The total sum-of-squares can now be split into **four** components.

$$SST = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (y_{ijk} - \bar{y})^2$$

r = days = 3; c = treatments = 4; n = no. of replicates = 2

$$SSA \text{ (between blocks or days sum-of-squares)} = nc \sum_{i=1}^r (\bar{y}_i - \bar{y})^2$$

$$SSB \text{ (between treatments or chelating agents sum-of-squares)} = nr \sum_{j=1}^c (\bar{y}_j - \bar{y})^2$$

$$SS(AB) \text{ (interaction sum-of-squares)} = n \sum_{i=1}^r \sum_{j=1}^c (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2$$

$$SS(\text{error}) = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij})^2$$

and $SST = SSA + SSB + SS(AB) + SSE$

The interaction term tests whether the performance of one factor depends on the level of the other e.g if the performance of the chelating agent was better on a given day.

In two-way ANOVA the calculations are quite tedious to carry out manually and are best done by computer.

General Two-Way ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed F
Main Effect				
A	SSA	a-1	$s1^2 = SSA/(a-1)$	$f1 = s1^2/s^2$
B	SSB	b-1	$s2^2 = SSB/(b-1)$	$f2 = s2^2/s^2$
Two-factor Interactions AB	SS(AB)	(a-1)(b-1)	$s3^2 = SS(AB)/(a-1)(b-1)$	$f3 = s3^2/s^2$
Error	SSE	ab(n-1)		
Total	SST	Abn-1		

As with one-way ANOVA the mean ss and F ratios are computed as outlined in the table above.

(Minitab) There are again two commands that can be used. The input of data for each command is the same. One column contains all the results. One column is used for codes to label which block (day) the result is from and one column for the labels for the treatments (chelating agents). These numbers are just labels and can be any integers but it is usual to use 1,2,3 ... The commands only differ in the form of the output.

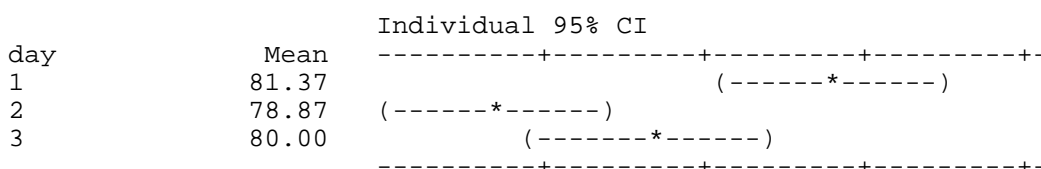
Two-way In this command the column where the results are is listed first and then the two columns where the factors are. By checking 'display' means' bar graphs indicating the degree of variability are given. These can be quite a useful pictorial representation.

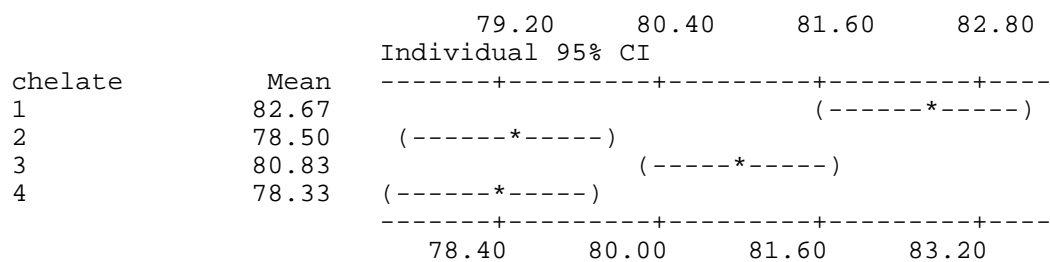
Balanced anova This command is the more general Minitab command, which can be used for Analysis of Variance with any number of factors. In this command the model is given - 'c1=c2|c3' means test for differences in factors stored in c2 and c3 plus any interactions ('c1 = c2 c3 c2*c3' would produce the same output). If the vertical bar between c2 and c3 was omitted the interaction effect would not be included.

Two-way Analysis of Variance

```

Analysis of Variance for %effic
Source      DF      SS      MS      F      P
day         2      25.08   12.54   10.03  0.003
chelate     3      76.83   25.61   20.49  0.000
Interaction 6      42.92    7.15    5.72  0.005
Error      12      15.00    1.25
Total      23     159.83
  
```





The results for this example show that both factors are significant and there is a significant interaction between days and chelating agents. What does this interaction mean? An inspection of the bar graphs shows that chelate C gives significantly higher % extractions than D but a closer look at the raw data shows that this difference is mainly due to differences on day 1. Thus differences between C and D depend on what day we are talking about.

1.8 Case Study: Comparison of Indicators

To further illustrate the use of two-way Analysis of Variance we can use the following data from a first year chemistry practical class. A class of 13 students carried out titrations of sodium carbonate with hydrochloric acid, using Methyl Red and Bromothymol Blue indicators. To test whether there is any significant difference in the end point determined using the different indicators we can use analysis of variance. The indicator difference is one source of variance but there is another (considerable) source of variance in the differences between students. In ANOVA testing the two sources of variances are separated and compared to the error (calculated from the differences between replicate determinations).

ANOVA testing needs balanced data (same number of replicates) so only the 'best' 3 titres of each student were used (i.e. the three that were closest together or most precise for each student).

The data and output for the two-way ANOVA are shown below. The results indicate ($p=0$) that differences between students and indicators are both highly significant. The interaction term is also significant - the different detection of endpoints using the indicators depends on the student.

The graph also give an idea of which students are the farthest away from the mean. Students 5 and 6 contribute the most variability. The data could be tested further to see if differences between students is still significant if we omit these two students' results. Note that the confidence intervals depicted in the graphs are pooled (i.e. average) C.I. Note also that in the case of indicator means the fact that the intervals do not overlap is further evidence of significant differences.

The two-way ANOVA has shown that despite large differences between students and a relatively small difference in the means of the two indicators (19.983 vs 20.238) we can still state confidently that this difference is significant.

	C1	C2	C3		C1	C2	C3
	titres	student	indicator		titres	student	indicator
1	19.60	1	1	40	20.15	1	2
2	19.58	1	1	41	20.1	1	2
3	19.59	1	1	42	20.18	1	2
4	19.94	2	1	43	20.23	2	2
5	19.81	2	1	44	20.19	2	2
6	19.73	2	1	45	20.19	2	2
7	19.7	3	1	46	20.3	3	2
8	19.69	3	1	47	20.29	3	2
9	19.7	3	1	48	20.28	3	2
10	19.5	4	1	49	20.2	4	2
11	19.5	4	1	50	20.2	4	2
12	19.5	4	1	51	20.2	4	2
13	20.92	5	1	52	20.87	5	2
14	20.73	5	1	53	20.85	5	2
15	20.61	5	1	54	20.99	5	2
16	20.52	6	1	55	21.61	6	2
17	19.7	6	1	56	21.51	6	2
18	19.87	6	1	57	21.32	6	2
19	19.61	7	1	58	20.1	7	2
20	19.71	7	1	59	20.25	7	2
21	19.67	7	1	60	20.11	7	2
22	19.66	8	1	61	19.8	8	2
23	19.58	8	1	62	19.89	8	2
24	19.7	8	1	63	19.84	8	2
25	20.59	9	1	64	20.14	9	2
26	20.58	9	1	65	20.08	9	2
27	20.62	9	1	66	20.11	9	2
28	19.99	10	1	67	20.22	10	2
29	20.02	10	1	68	20.38	10	2
30	19.97	10	1	69	20.39	10	2
31	20.22	11	1	70	19.62	11	2

32	20.26	11	1	71	19.88	11	2
33	20.3	11	1	72	19.81	11	2
34	20.1	12	1	73	19.2	12	2
35	20.71	12	1	74	19.17	12	2
36	20.2	12	1	75	19.8	12	2
37	19.85	13	1	76	20.22	13	2
38	19.92	13	1	77	20.3	13	2
39	19.87	13	1	78	20.32	13	2

Two-way Analysis of Variance

Analysis of Variance for titres

Source	DF	SS	MS	F	P
student	12	8.3281	0.6940	33.64	0.000
indic	1	1.2744	1.2744	61.78	0.000
Interaction	12	6.7361	0.5613	27.21	0.000
Error	52	1.0727	0.0206		
Total	77	17.4113			

