CHAPTER – 4 STATISTICS

Dr. S. RENGARAJ., Ph.D

Definition of Statistics:

1. **A collection of data or numbers.**

2. **Set of mathematical tools used to describe and make judgments about data.**

3. **Logic which makes use of mathematics in the science of collecting, analyzing and interpreting data for the purpose of making decision.**

Type of statistics we will talk about in this class has important assumption associated with it: Experimental variation in the population from which samples are drawn has a normal (Gaussian, Bell-shaped) distribution.

Number of data: n

Mean:

Average or Arithmetic Mean.

The arithmetic mean, $\bar{x}$ - also called the average – is the sum of the measured values divided by *n*, the number of measurements:

The mean gives the center of the distribution.

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

Where:

$\bar{x}$ - **is the mean value.**

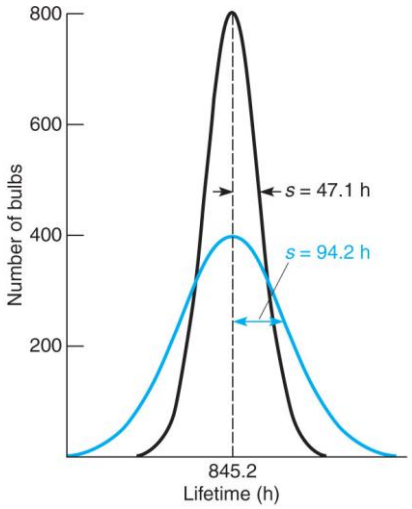
x_i - **is the individual value.**

$\sum$ - **is the summation.**

n - **is the number of observation.**

Standard Deviation:

The standard deviation, S, measures how closely the data are clustered about mean.



The smaller the standard deviation, the more closely the data are clustered about the mean.

The mean gives the center of the distribution. The standard deviation measures the width of the distribution.

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

where

S	-	**is the standard deviation.**
x_i	-	**is the individual value.**
$\bar{x}$	-	**is the mean value.**
Σ	-	**is the summation.**
S^2	-	**is the variance.**
n	-	**is the number of observations.**

Variance:

Used in many other statistical calculations and tests.

The square of the standard deviation is called the variance.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

where

S^2 - **is the variance.**

x_i - **is the individual value.**

$\bar{x}$ - **is the mean value.**

$\sum$ - **is the summation.**

n - **is the number of observation.**

Or

Variance = S^2

Relative Standard Deviation or Coefficient of variation:

The standard deviation expressed as a percentage of the mean value is called the relative standard deviation or the coefficient of variation.

RSD or CV = $\frac{S}{\bar{x}} \times 100\%$

Where

S - **is the standard deviation.**

$\bar{x}$ - **is the mean value.**

Range or spread (w):

Range or spread (w) = Largest value – smallest value.

Gap:

Gap = Difference between questionable data point and its nearest neighbor.

Standard Error:

Tells us that standard deviation of set of samples should decrease if we take more measurements

Standard error = $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

Average deviation:

Another way to express degree of scatter or uncertainty in data. Not as statistically meaningful as standard deviation, but useful for small samples.

$$\bar{d} = \frac{\sum_i (|x_i - \bar{x}|)}{n}$$

Relative average deviation (RAD)

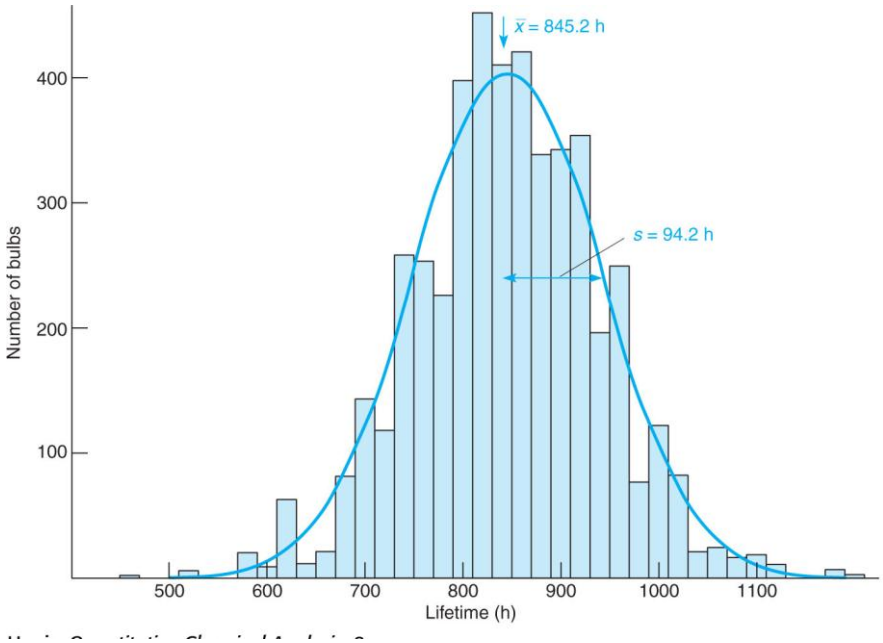
$$RAD = \left(\frac{\bar{d}}{\bar{x}}\right)100 \quad (as\ percentage)$$

$$RAD = \left(\frac{\bar{d}}{\bar{x}}\right)1000 \quad (as\ parts\ per\ thousand,\ ppt)$$

4-1 Gaussian Distribution

Normal distribution or Gaussian distribution:

The number of experiment is repeated, the more closely the results approach and ideal smooth curve called the Gausian distribution.



* **Infinite members of group – Population.**
* **Characterize population by taking samples.**
* **The larger the number of samples, the closer the distribution becomes to normal.**
* **Equation of normal distribution:**

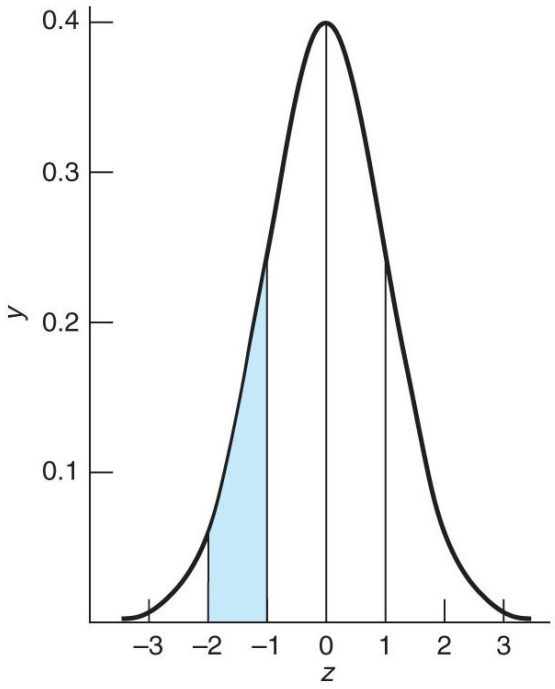
$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Where:

σ = population standard deviation, S

μ = population mean, $\bar{x}$

For a finite set of data, we approximate *μ* by $\bar{x}$ and *σ* by *S*.



[A Gaussian curve in which μ = 0 and σ = 1. A Gaussian curve whose area is unity is called a normal error curve. The abscissa z = (x- μ)/ σ is the distance away the mean, mean, measured in units of the standard deviation. When z = 2, we are two standard deviations away from the mean.]
[When z = +1, x is one standard deviation above the mean. When z = -2, x is two standard deviation below the mean.]

It is useful to express deviation from the mean value in multiples, z, of the standard deviation. That is we transform x into z given by

$$z = \frac{x - \mu}{\sigma} \approx \frac{x - \bar{x}}{S}$$

Estimate of mean value of population = μ

Estimate of mean value of samples = $\bar{x}$

(As the number of measurements increases, $\bar{x}$ approaches μ, if there is no systematic error.

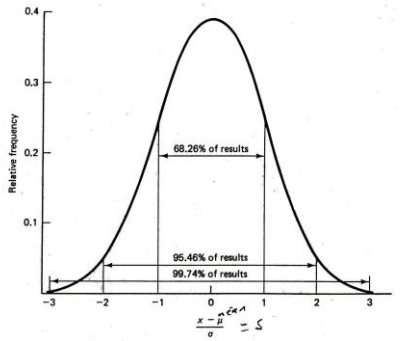


Figure 2.2 Normal distribution curve; relative frequencies of deviations from the mean for a normally distributed infinite population; deviations ($x - \mu$) are in units of σ.

Mean = $\bar{x} = \dfrac{\sum_{i=1}^{N} x_i}{n}$

Degree of scatter (measure of central tendency) of population is quantified by calculating the *standard deviation*.

Std. dev. of population = σ
Std. dev. of sample = s

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

Characterize sample by calculating $\bar{x} \pm S$

Standard deviation and the normal distribution:

- **Standard deviation defines the shape of the normal distribution (particularly width).**
- **Larger std. dev. means more scatter about the mean, worse precision.**

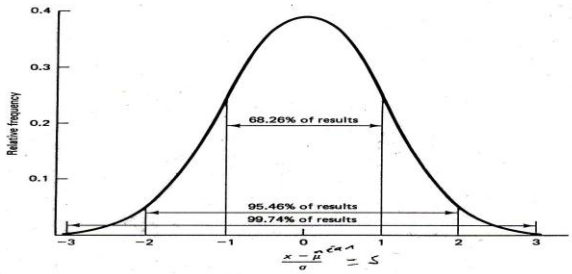


Figure 2.2 Normal distribution curve; relative frequencies of deviations from the mean for a normally distributed infinite population; deviations $(x - \mu)$ are in units of σ.

- **Smaller std. dev. means less scatter about the mean, better precision.**

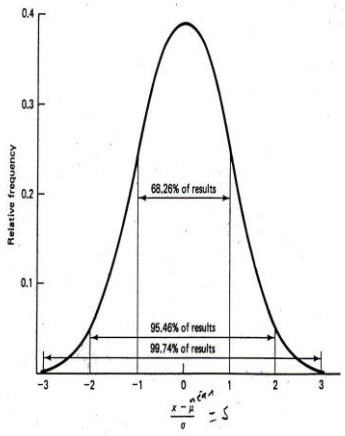


Figure 2.2 Normal distribution curve; relative frequencies of deviations from the mean for a normally distributed infinite population; deviations $(x - \mu)$ are in units of σ.

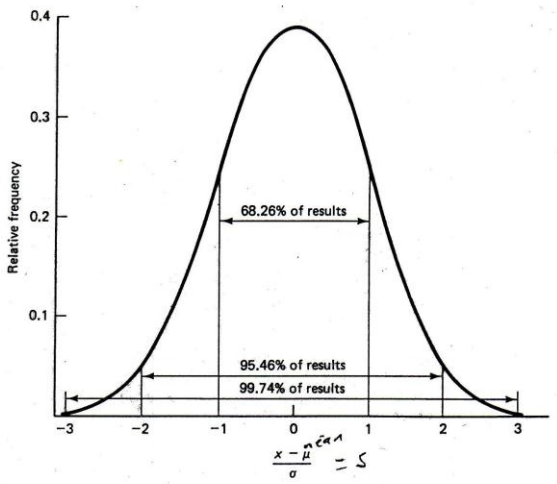


Figure 2.2 Normal distribution curve; relative frequencies of deviations from the mean for a normally distributed infinite population; deviations $(x - \mu)$ are in units of σ.

There is a well-defined relationship between the std. dev. of a population and the normal distribution of the population:

$\mu \pm 1\sigma$ encompasses 68.3 % of measurements
$\mu \pm 2\sigma$ encompasses 95.5% of measurements
$\mu \pm 3\sigma$ encompasses 99.7% of measurements

(May also consider these percentages of area under the curve)

TABLE 4–1 **Ordinate and area for the normal (Gaussian) error curve, $y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$**

$\lvert z\rvert$[a]	y	Area[b]	$\lvert z\rvert$	y	Area	$\lvert z\rvert$	y	Area
0.0	0.398 9	0.000 0	1.4	0.149 7	0.419 2	2.8	0.007 9	0.497 4
0.1	0.397 0	0.039 8	1.5	0.129 5	0.433 2	2.9	0.006 0	0.498 1
0.2	0.391 0	0.079 3	1.6	0.110 9	0.445 2	3.0	0.004 4	0.498 650
0.3	0.381 4	0.117 9	1.7	0.094 1	0.455 4	3.1	0.003 3	0.499 032
0.4	0.368 3	0.155 4	1.8	0.079 0	0.464 1	3.2	0.002 4	0.499 313
0.5	0.352 1	0.191 5	1.9	0.065 6	0.471 3	3.3	0.001 7	0.499 517
0.6	0.333 2	0.225 8	2.0	0.054 0	0.477 3	3.4	0.001 2	0.499 663
0.7	0.312 3	0.258 0	2.1	0.044 0	0.482 1	3.5	0.000 9	0.499 767
0.8	0.289 7	0.288 1	2.2	0.035 5	0.486 1	3.6	0.000 6	0.499 841
0.9	0.266 1	0.315 9	2.3	0.028 3	0.489 3	3.7	0.000 4	0.499 904
1.0	0.242 0	0.341 3	2.4	0.022 4	0.491 8	3.8	0.000 3	0.499 928
1.1	0.217 9	0.364 3	2.5	0.017 5	0.493 8	3.9	0.000 2	0.499 952
1.2	0.194 2	0.384 9	2.6	0.013 6	0.495 3	4.0	0.000 1	0.499 968
1.3	0.171 4	0.403 2	2.7	0.010 4	0.496 5	∞	0	0.5

a. $z = (x - \mu)/\sigma$.

b. The area refers to the area between $z = 0$ and $z =$ the value in the table. Thus the area from $z = 0$ to $z = 1.4$ is 0.419 2. The area from $z = -0.7$ to $z = 0$ is the same as from $z = 0$ to $z = 0.7$. The area from $z = -0.5$ to $z = +0.3$ is $(0.191\ 5 + 0.117\ 9) = 0.309\ 4$. The total area between $z = -\infty$ and $z = +\infty$ is unity.

Some useful Statistical Tests:

- **To characterize or make judgments about data.**

- **Tests that use the *Student's t distribution***
  - **Confidence intervals.**
  - **Comparing a measured result with a "known" value.**
  - **Comparing replicate measurements (comparison of means of two sets of data).**

4-2 Confidence interval:

From a limited number of measurements (n), we cannot find the true population mean, *μ*, or the true standard deviation, σ. What we determine are $\bar{x}$ and S, the sample mean and sample standard deviation.

Quantifies how far the true mean (μ) lies from the measured mean, Uses the mean and standard deviation of the sample.

The confidence interval is computed from the equation

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

Where

S	-	**is the measured standard deviation.**
$\bar{x}$	-	**is the mean value.**
n	-	**is the number of observations.**
t	-	**is the student's 't' taken from the table.**

Degrees of freedom (df) = *n* - 1 for the CI.

4-3 Comparison of mean with students t:

If you make two sets of measurements of the same quantity, the mean value from one set will generally not be equal to the mean value from the other set because of small, random variations in the measurements.

Therefore, we use a t-test to compare one mean value with another to decide whether there is a statistically significant difference between the two.

Case I: Comparing a measured result with a "known" value:

We measure a quantity several times, obtaining an average value and standard deviation. We need to compare our answer with an accepted answer. The average is not exactly the same as the accepted answer. Does our measured answer agree with the accepted answer "within experimental error?.

95% confidence interval = $\bar{x} \pm \frac{ts}{\sqrt{n}}$

$$t_{calculated} = \frac{|\bar{x} - \mu|}{S}\sqrt{n}$$

Where

S	-	**is the measured standard deviation.**
$\bar{x}$	-	**is the mean value.**
μ	-	**known value.**
n	-	**is the number of observations.**

'$t_{calculated} < t_{table}$ => no significance difference.
$t_{calculated} > t_{table}$ => There is significance difference.

df = n -1 for this test

Case II: **Comparing replicate measurements or comparing means of two sets of data:**

We measure a quantity multiple times by two different methods that give two different answers, each with its own standard deviation. Do the results agree with each other "within experimental error"?

Example: Given the same sample analyzed by two different methods, do the two methods give the "same" result?

$$t_{calculated} = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{S_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$S_{pooled} = \sqrt{\frac{\sum_{set1}(x_i - \overline{x_1})^2 + \sum_{set2}(x_i - \overline{x_2})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)}{n_1 + n_2 - 2}}$$

Will compare t_{calc} to tabulated value of *t* at appropriate df and CL.

df = $n_1 + n_2 - 2$ for this test

Case III: Paired t test for Comparing individual differences:

Sample A is measured once by method 1 and once by method 2; the two measurements do not give exactly the same result. Then a different sample, designated B, is measured once by method 1 and once by method 2; and, again, the results are not exactly equal. The procedure is repeated for n different samples. Do the two methods agree with each other "within experimental error"?

$$t_{calculated} = \frac{\bar{d}}{S_d}\sqrt{n}$$

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

Where

$d_i = x_{1i} - x_{2i}$

$\bar{d} = \sum \frac{d_i}{n}$

$S_d^2 = \frac{\sum(d_i - \bar{d})^2}{n-1}$

df = n - 1 for this test

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	63.656	127.321	636.578
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

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.

Values of t in this table apply to two-tailed tests illustrated in Figure 4-9a. The 95% confidence level specifies the regions containing 2.5% of the area in each wing of the curve. For a one-tailed test, we use values of t listed for 90% confidence. Each wing outside of t for 90% confidence contains 5% of the area of the curve.

Harris, *Quantitative Chemical Analysis*, 8e

4-4 Comparison of standard deviation with the F Test:

The *F* test tells us whether two standard deviations are "significantly" different from each other.

***F* is the quotient of the squares of the standard deviations:**

$$F_{Calculated} = \frac{S_1^2}{S_2^2}$$

We always put the larger standard deviation in the numerator so that $F \geq 1$.

We test the hypothesis that $S_1 > S_2$ by using the one-tailed F test in Table 4-4.

If $F_{calculated} > F_{Table}$, then the difference is significant.

- **Used to determine if std. devs. are significantly different before application of *t*-test to compare replicate measurements or compare means of two sets of data.**
- **Also used as a simple general test to compare the precision (as measured by the std. devs.) of two sets of data.**
- **Uses F distribution.**

Will compute F_{calc} and compare to F_{table}.

$$F_{calc} = \frac{s_1^2}{s_2^2} \quad where \quad s_1 > s_2$$

Degrees of Freedom = $n_1 - 1$ and $n_2 - 1$ for this test.

Choose confidence level (95% is a typical CL).

Note that the F-test can be used to simply test whether or not two sets of data have statistically similar precisions or not.

TABLE 4-4 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

Critical values of F for a one-tailed test of the hypothesis that $s_1 > s_2$. There is a 5% probability of observing F above the tabulated value.

You can compute F for a chosen level of confidence with the Excel function FINV(probability,deg_freedom1,deg_freedom2). The statement "=FINV(0.05,7,6)" reproduces the value F = 4.21 in this table. The statement "=FINV(0.1,7,6)" gives F = 3.01 for 90% confidence.

Harris, *Quantitative Chemical Analysis*, 8e

Q-test for Bad data:

Evaluating questionable data points using the Q-test

- **Need a way to test questionable data points (outliers) in an unbiased way.**
- **Q-test is a common method to do this.**
- **Requires 4 or more data points to apply.**

Calculate Q_{calc} and compare to Q_{table}

$$Q_{Calculated} = \frac{gap}{Range}$$

Gap = (difference between questionable data pt. and its nearest neighbor)
Range = (largest data point – smallest data point)

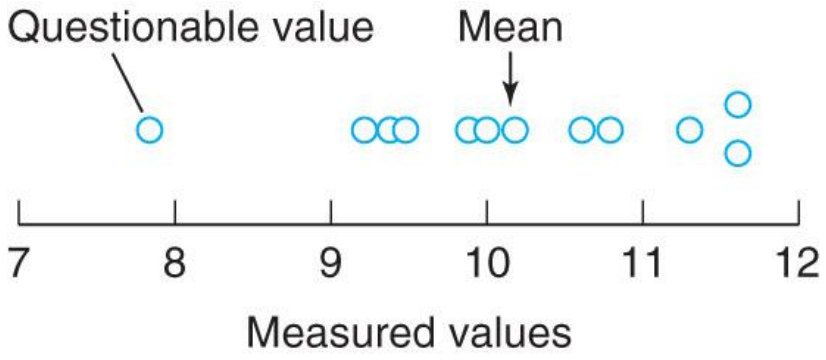
If $Q_{calc} < Q_{table}$, do not reject questionable data point at stated CL.
If $Q_{calc} \geq Q_{table}$, reject questionable data point at stated CL.

Rejection Quotient, *Q*, at Different Confidence Limits[a]

No. of Observations	*Confidence level*		
	Q_{90}	Q_{95}	Q_{99}
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568
15	0.338	0.384	0.475
20	0.300	0.342	0.425
25	0.277	0.317	0.393
30	0.260	0.298	0.372

[a]Adapted from D. B. Rorabacher, *Anal. Chem.*, **63** (1991) 139.

4-6 Grubbs test for an Outlier:



A datum that is far from other points is called an outlier.

$$G_{Calculated} = \frac{\left| questionable value - \bar{x} \right|}{S}$$

Where the numerator is the absolute value of the difference between the suspected outlier and the mean value.

If $G_{calculated}$ is greater than G in Table 4-5, the questionable point should be discarded.

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.*