Introduction to EXP2-1

EXP2-1 Same or Different?

Semester B, Weeks 1-3

The **aim** of EXP2-1 is to develop tools and skills, in order to:

- Present and report data in the correct way

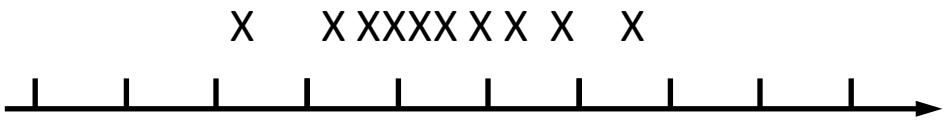
  (units, significant figures, error, and others)
- Determine whether the data are statistically significant

Introduction to EXP2-1 Part 1

Part 1 Random variations

When we measure a value repeatedly, we do not get exactly the same result each time

Typically, we see random variation about some value



Repeat the measurement to reveal random variation

Random variation

Some definitions

Variables may be

Discrete or categorical: only have specific values

e.g. Women or Men

number of legs

polymer students or materials students

Continuous or numerical: can have any value

e.g. most things we measure: length, temperature, mass, strength, stiffness etc.

Random variation

Some definitions

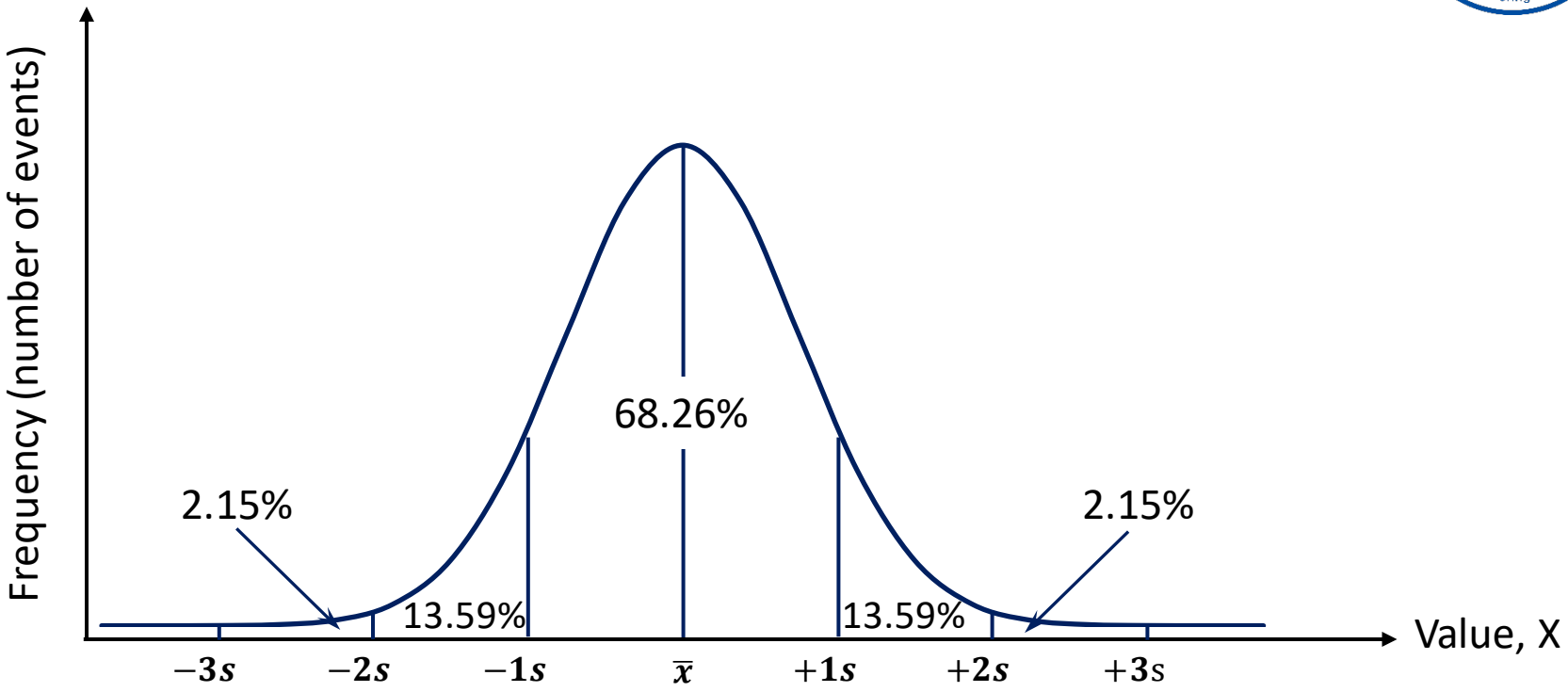
Data: individual measured values (observations) (e.g. Student's X height)

Sample: a number of similar measured values (e.g. QMES students' height)

Population: hypothetical set of all results (e.g. Chinese students' height)

A population of random variation follows the **"Normal Distribution".**

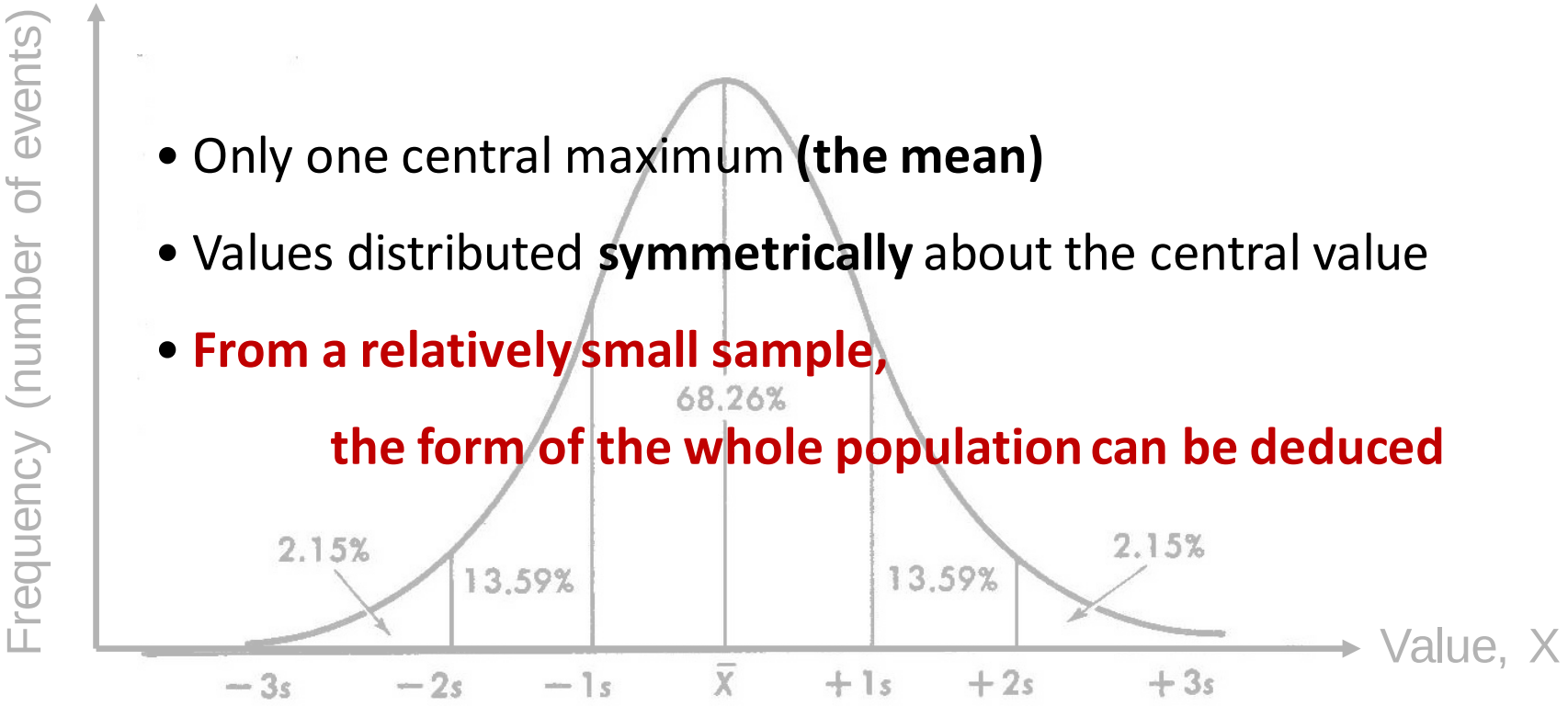
The Normal Distribution



Expresses the probability of a value lying within a certain distance of the mean

The Normal Distribution

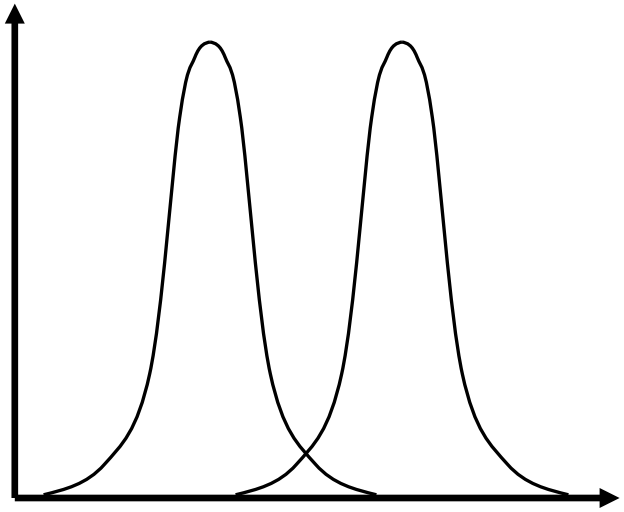
Properties of the Normal Distribution



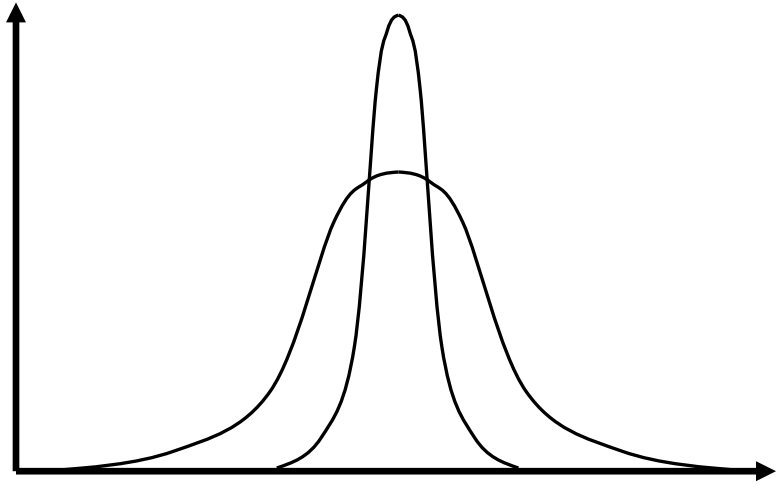
- Only one central maximum **(the mean)**
- Values distributed **symmetrically** about the central value
- **From a relatively small sample,**

  the form of the whole population can be deduced

Mean and dispersion



Same dispersion different mean



Same mean different dispersion

Measures of central tendency

- Mean (average)
- Mode
- Median

Measures of central tendency

Mean

- For the population: $\mu = \frac{\sum X_i}{N}$
- For the sample: $\bar{X} = \frac{\sum X_i}{n}$

Mode

- Most commonly occurring value in a set of numbers
- Usually used when data grouped together in classes – is then sensitive to class interval.
- Can have 1 value (unimodal) or >1 values (e.g. bimodal, trimodal)

Measures of central tendency

Median

- Defined by position **(n+1)/2** along dataset

Example 1: 1 5 **6** 8 9

Example 2: 200 200 230 250 **280** 300 350 400 480

Example 3: 200 200 230 250 **280 300** 350 400 480 500

Median = (280 + 300) / 2 = **290**

Characterising variation

Characterising location

Central tendency:

	Sample	Population
Mean or average value	$\overline{X} = \frac{\sum X}{n}$	μ

n = number of events (or number of measurements)

Characterising variation

Characterising spread

Dispersion:

	Sample	Population
Range	$X_{max} - X_{min}$	$+\infty$ to $-\infty$
Standard deviation, (SD or St Dev)	$S = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum(X - \bar{X})^2}{N}}$

Characterising variation

Standard deviation, SD

Within

1 SD = 68.26% of population

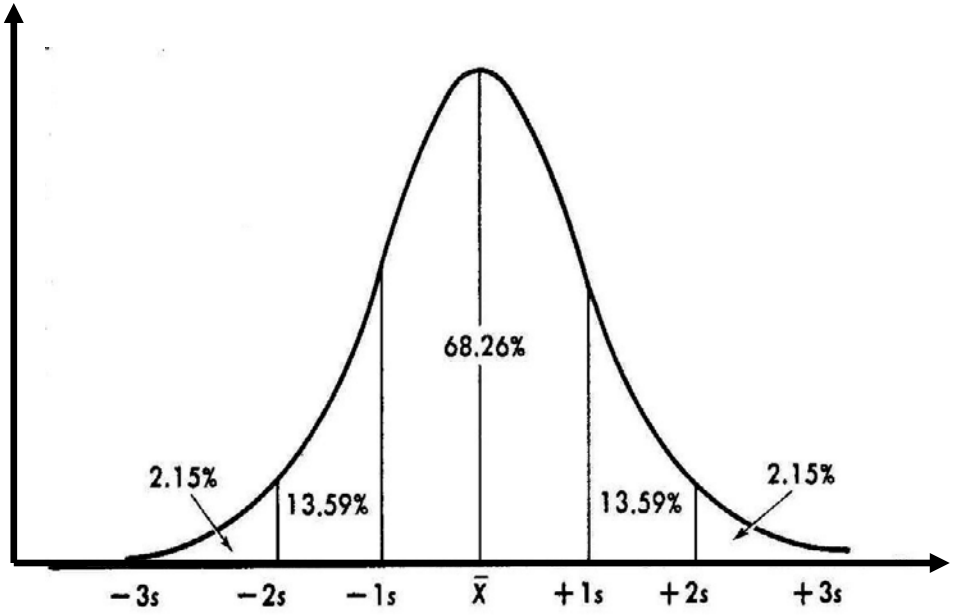
2 SD = 95.46% of population

3 SD = 99.73% of population

…

Out of 6 SD

≈ 1/1million of population



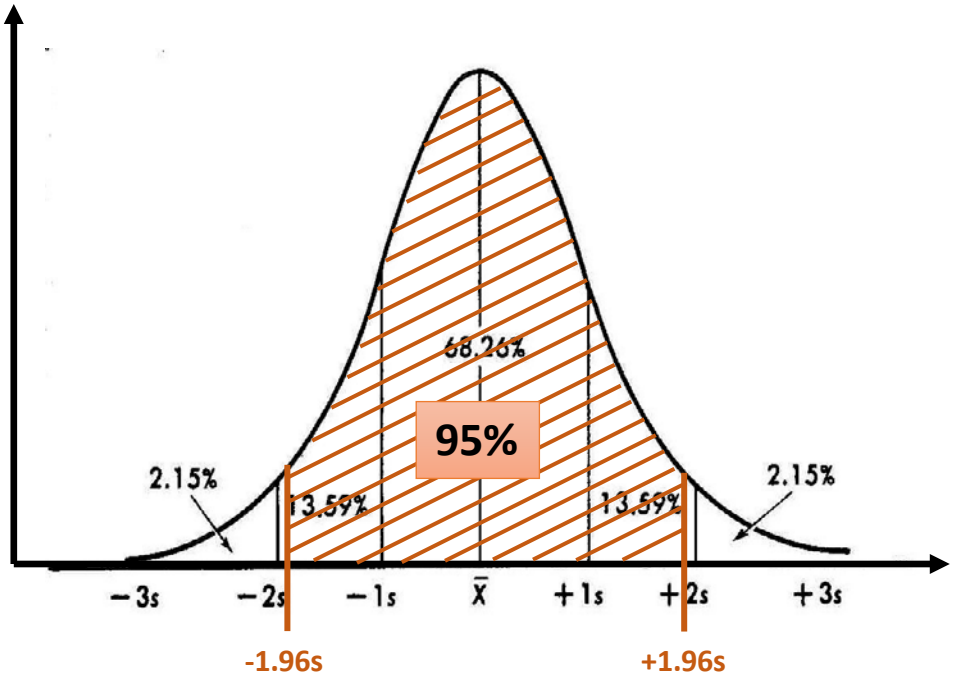
Expresses the probability of a value lying within a certain distance of the mean.

Characterising variation

Confidence intervals, CI

Probability of x% of values occurring within certain limit of the mean

i.e. How confident are we that a value will be found in this range?



Example:

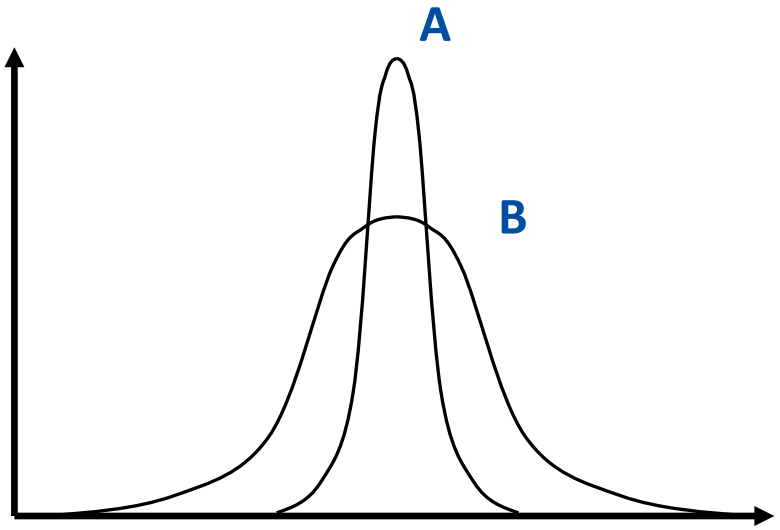
95% confidence interval = ± 1.96 SD of mean

99% confidence interval = ± 2.58 SD of mean

Characterising variation

Coefficient of variation, CV

$$C.V. = \frac{St\ Dev}{Mean}$$



Gives a measure of the width

or sharpness of the distribution.

e.g., CV(A) < CV(B)

$$Usually\ expressed\ as\ a\ percentage\ \ CV\% = \frac{S}{\bar{X}} * 100$$

Reporting data

In science – data do not have a single value,

we also need to show *how well we know the value*

Location ± **spread** (n = number of measurements)

Mean ± **St Dev, n**

Mean ± ***x*% confidence intervals, n**

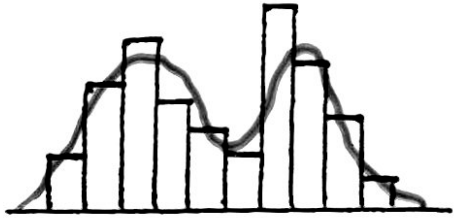
Mean + CV, n

} Use 1 of these

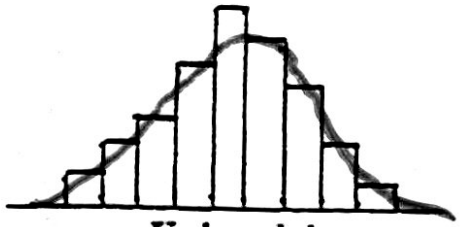
Always state what it is that you have quoted

There are other ways to express variation

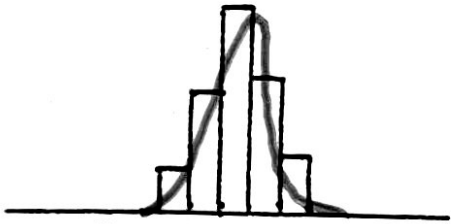
Other forms of distribution



Bimodal

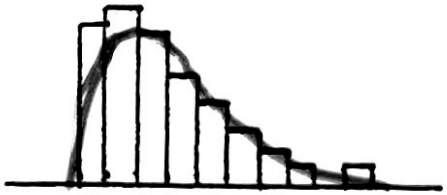


Unimodal

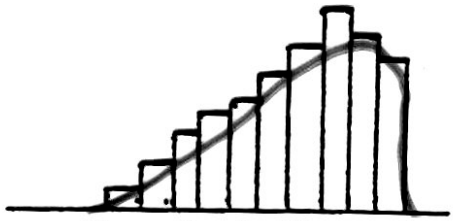


Small Variability

Large Variability



Positively Skewed



Negatively Skewed

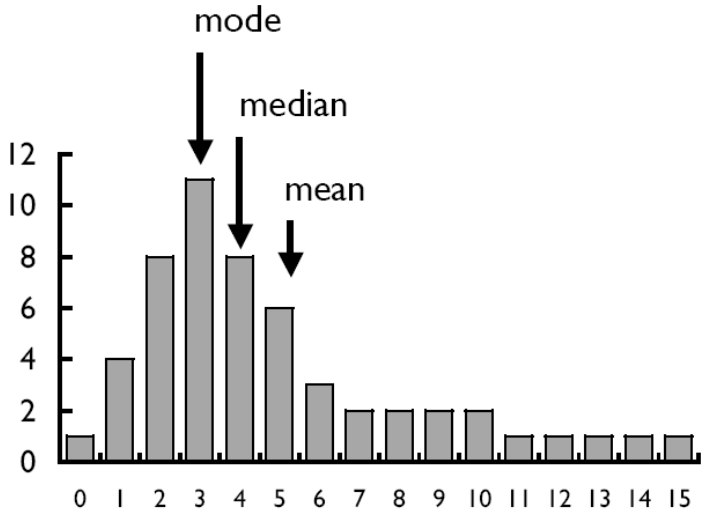
Other forms of distribution

Central tendency:

Mean = average value

Median = middle value

Mode = most frequently occurring value



Dispersion:

Histograms – divide the data into categories or 'bins'

Determine the number of bins:

use $\sqrt{n}$ bins <u>or</u> *Sturges rule* $1 + 3.3 \log n$ bins

(where n = the total number of observation in the dataset)

Sampling distribution

- **Population** - the entire group
- **Sample** - the specific group

- Probability samples do not guarantee '**correct**' answers in terms of their estimate of the underlying **population** value

- **Each sample varies** and so each sample's estimate will be different (hopefully not by much)

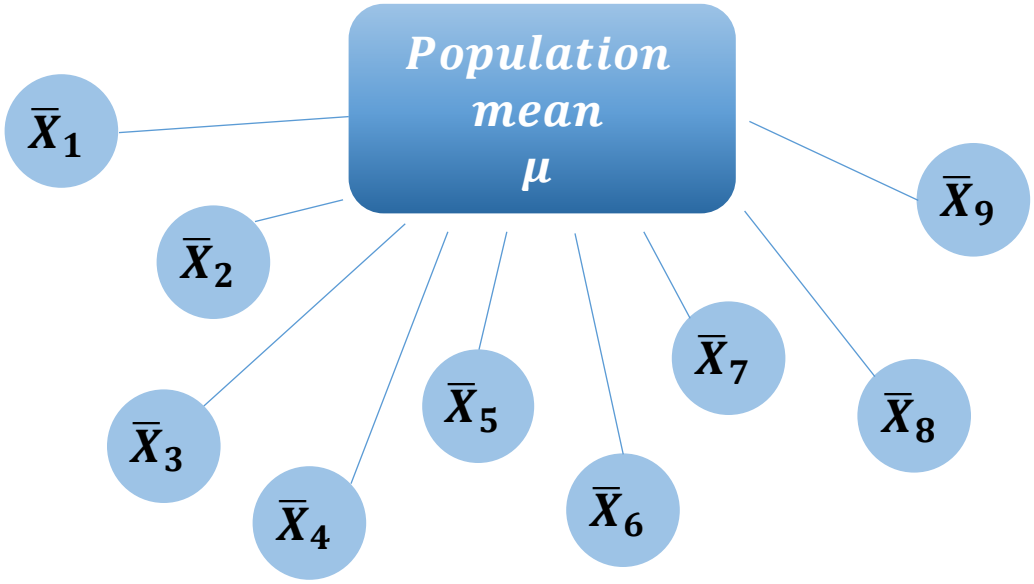
- Still, one sample is just one of many **possible samples** from the population

Sampling distribution

- Question: If we take only one sample, can we **estimate** how precise its sample estimates are in terms of estimating the **true** (but unknown) **population** value?
  Yes - if we have **random** samples!

- And why is it important?
  Because we can then **generalise** from our **sample** to the **population**!
  And - if provided we sample well - we only need to take relatively **few cases** in order to calculate fairly precise estimates for much **bigger populations**.

Sample means



Samples: 1, 2, 3, …

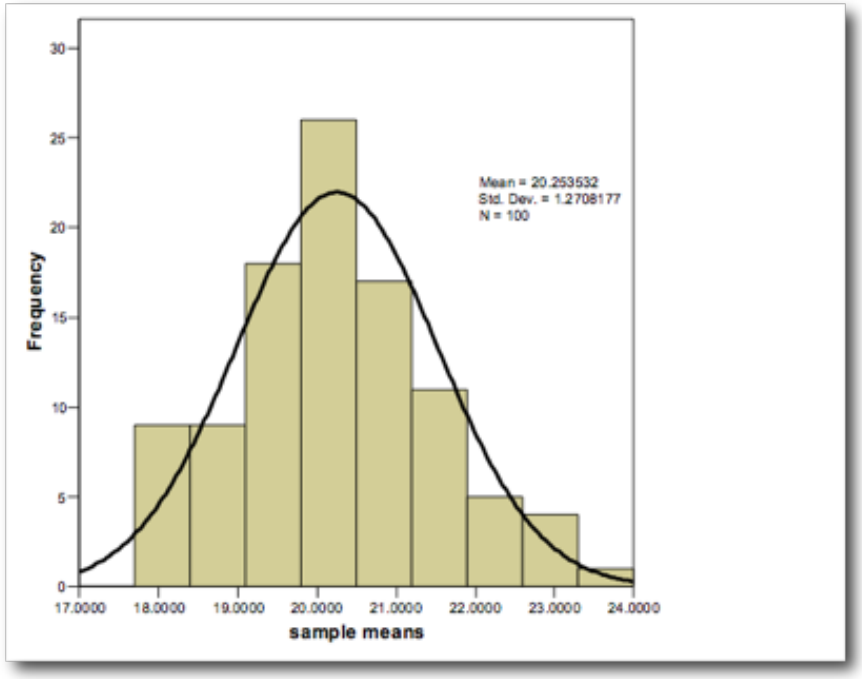
Sample means: $\bar{X}_1, \bar{X}_2, \bar{X}_3$…

- **Sample means** are random variations
- Sample means centered on population mean, with enough samples, average of all sample means = true population mean

Standard Error of the Mean

Sampling distribution is the plot of sample means**.**

Standard error of mean = standard deviation of the sampling distribution



Standard error (SE) of mean becomes important:

If we know what SE is then we can assess how **likely** it is that a given **sample** is **representative** of the '**true**' population value.

Central Limit Theroem

If our sample is relatively large (n > 30),

then the **Central Limit Theorem** holds true.

This tells us that the sampling distribution will have:

- **Normal** distribution
- Sample mean equals the population mean $\bar{X} = \mu$
- Standard deviation of sampling distribution $\mathrm{SE} = \frac{\sigma}{\sqrt{\mathrm{n}}}$
- Known properties (e.g. **95%** of the sample means will lie within 1.96 standard errors of the population mean), 95% C.I.= $\pm$ 1.96SE
- If n > 30, then this holds even if the target population is **not** normally distributed!

Standard Error of the Mean

Standard error (SE) of a sample estimate:

We can use **Central Limit Theorem** to estimate the **accuracy** of the sample estimate of the mean

SE (standard deviation of sampling distribution): $\boldsymbol{\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}}$

SE indicated the likely **spread** of the **sample means** around the **population mean**

- a measure of sampling error and an indication of how **likely** it is that our sample estimate is close to the **true** underlying **population** value.

And when n > 30 we can use the equation above to estimate it.

Standard Error of the Mean

Standard error of the sampling distribution:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$SD\ of\ population\ \sigma = \sqrt{\frac{\sum(X - \bar{X})^2}{N}}$$

$$\sigma_{\bar{X}} = \frac{S}{\sqrt{n}}$$

$$SD\ of\ sample\ S = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$$

Implications:

- The **larger the sample** size (*n*), the **smaller the error** in the sample is likely to be.
- The **larger the standard deviation** in the sample (*S*), the **larger the error** in the sample is likely to be.

SE & Confidence interval

Standard error of estimate allows us to calculate a **confidence interval** around our sample estimate.

Confidence interval = a **range** within which we are reasonably sure the **real population mean lies**.

What does the confidence interval **mean**?

For a 95% confidence interval, think that if you collected 100 samples then the true population mean would lie within the confidence intervals of 95 of those samples.

SE & C.I. of sampling distribution

When n > 30 the CLT holds and confidence intervals for sample means:

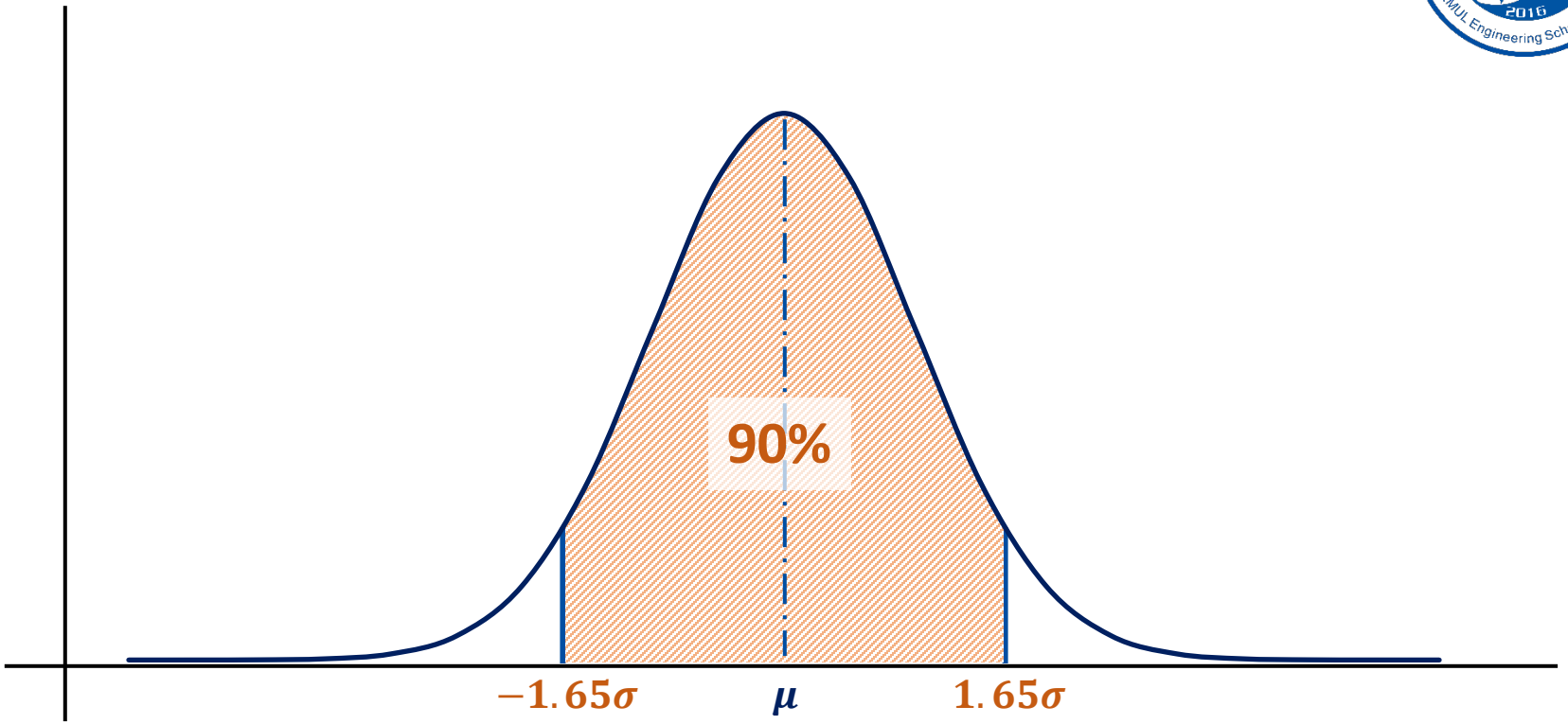
$$\boldsymbol{Confidence\ interval} = \overline{X} \pm Z \cdot \sigma_{\overline{X}}$$

sample estimate of the mean

Standard error (standard deviation of sample / square root of n)

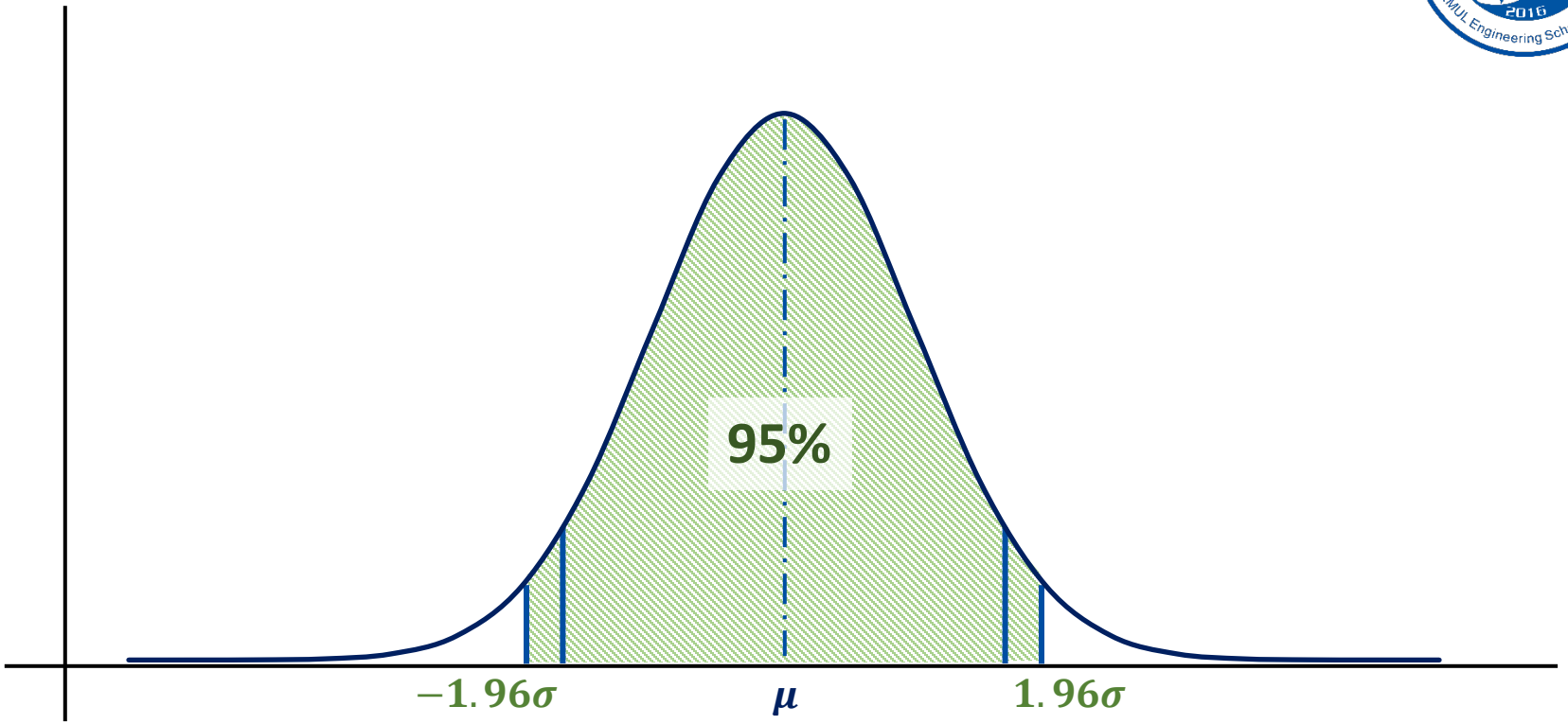
Z is a z-value taken from the normal curve. Value of Z depends on how much error we are willing to accept.

Confidence Intervals



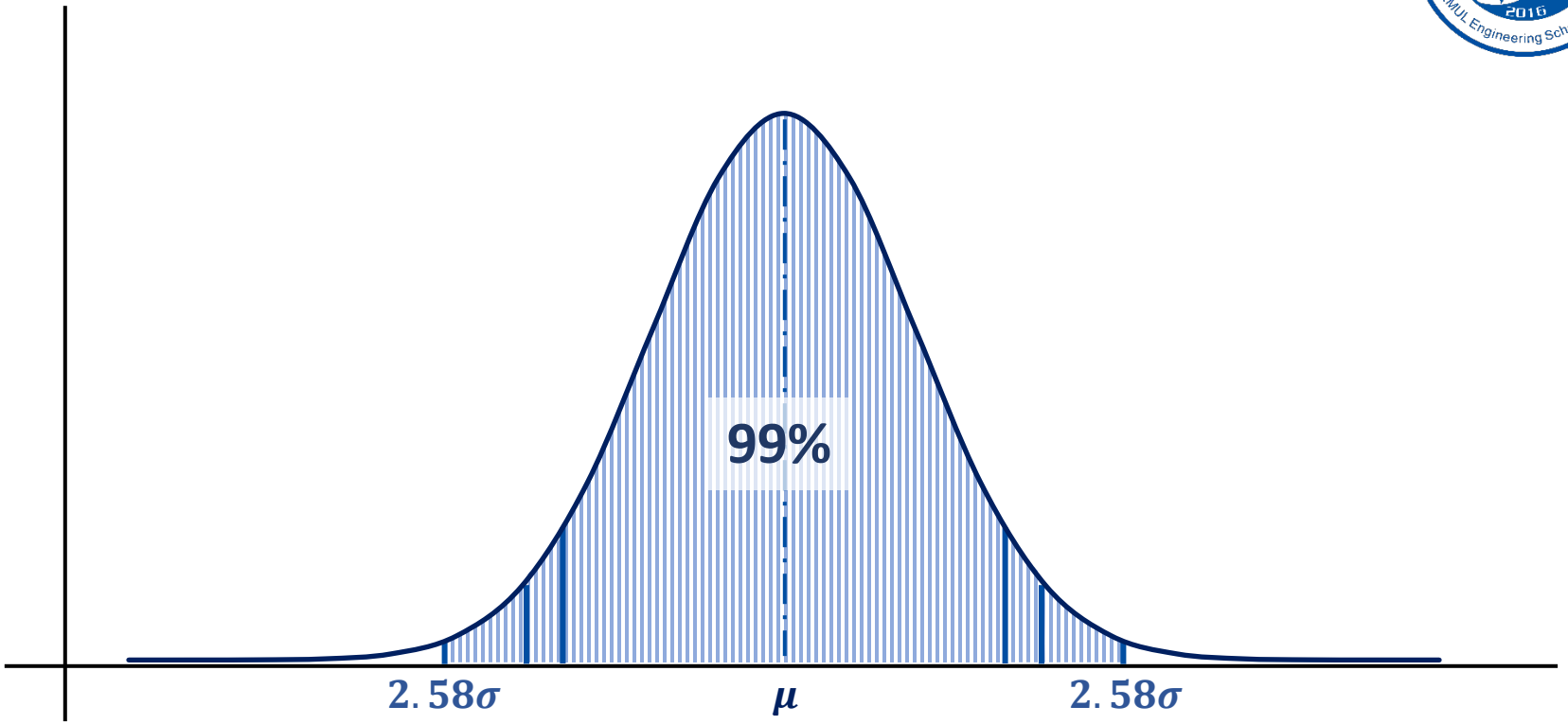
Confidence interval	90%
z	1.65

Confidence Intervals



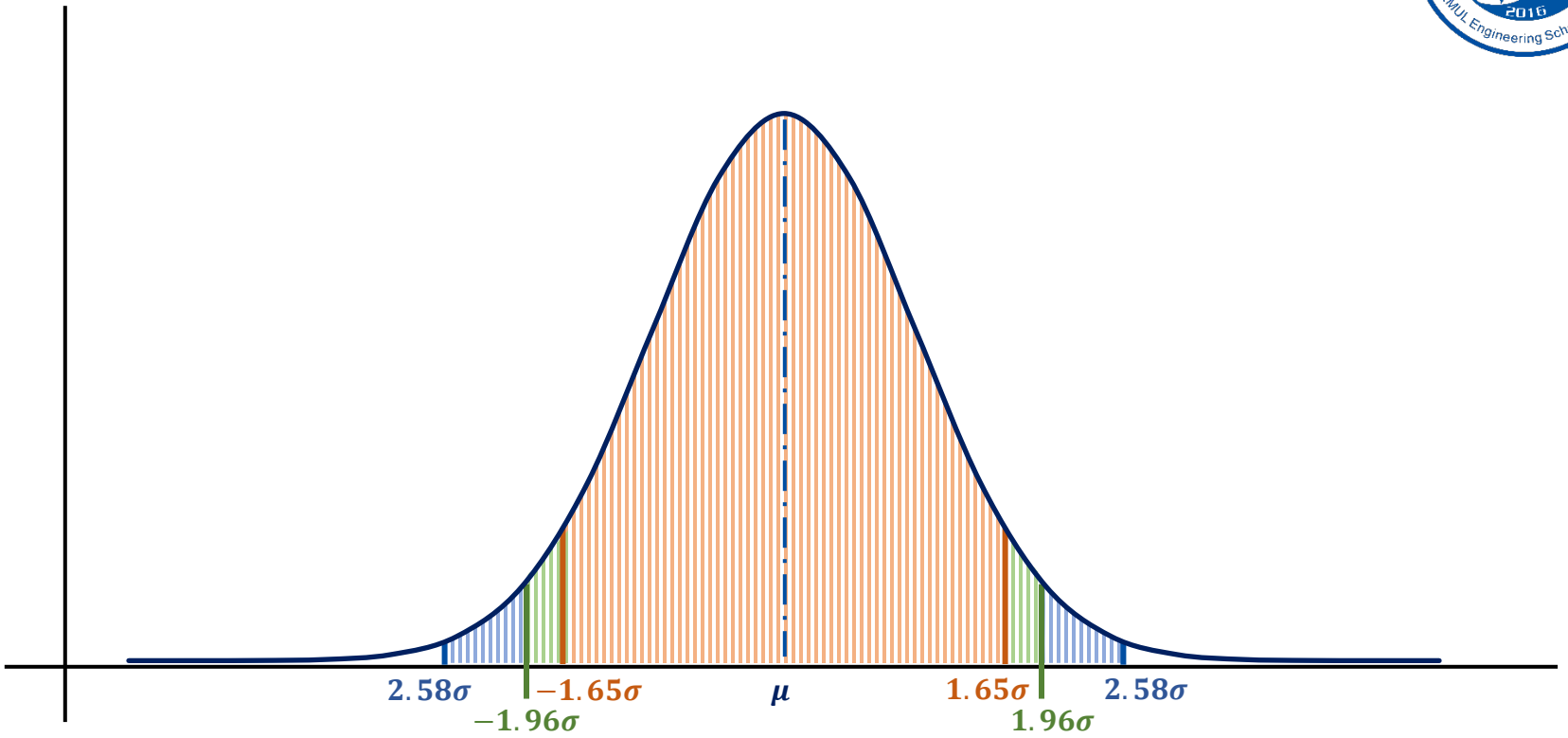
Confidence interval	90%	95%
z	1.65	1.96

Confidence Intervals



Confidence interval	90%	95%	99%
z	1.65	1.96	2.58

Confidence Intervals



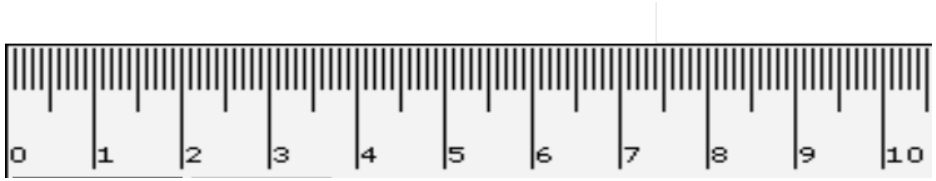
Confidence interval	90%	95%	99%
z	1.65	1.96	**2.58**

Summary of random distribution

- **Random variation** follows the **normal distribution.**
- **St Dev** expresses the dispersion about the mean value.
- **SE** (Standard Error of the Mean) expresses how far a sample mean is from the "true" value (the population mean).
- **Confidence intervals** express the probability of any value being within a certain distance of the mean.
- **Coefficient of variation** expresses the width of the distribution.
- **Histograms** can be useful in showing distributed values.

Significance Figures

- Measure the same object, the output significant digits is based on the tools. How accurate can it measure?



l = **5.6** mm Unit： mm

Certain value

Vs.

Estimated value

Significant digits

l = **5.678** mm Unit： mm

Tasks

- Check your group assignments on QM+
- Review the ‘Introduction’ and follow the ‘Instructions’
- Download the raw data files from QM+
- Follow 7-step plan, hold your GMM to solve the problem
- Bring your questions to be discussed in class

EXP2-1 Schedule – M Groups

Date:	Time:	Place:	Activity:
Monday 26th	14:00-15:40	A310	Introduction to QXU5017 Introduction to EXP2-1 Part1
Wednesday 28th	16:00-17:40	B1-202	Group Meeting 1 (Review and check the raw data)
Thursday 29th	16:00-17:40	B1-202	Group Meeting 2 (Write Report Section 1-4)
Monday 4th	14:00-15:40	A310	Introduction to EXP2-1 Part2
Wednesday 6th	16:00-17:40	B1-202	Group Meeting 3 (Different sample comparison)
Thursday 7th	16:00-17:40	B1-202	Group Meeting 4 (Complete report Section 5-8)
Monday 11th	14:00-15:40	A310	Report submission

EXP2-1 Schedule – P Groups

Date:	Time:	Place:	Activity:
Monday 26th	16:00-17:40	A310	Introduction to QXU5017 Introduction to EXP2-1 Part1
Wednesday 28st	14:00-15:40	B1-302	Group Meeting 1 (Review and check the raw data)
Thursday 29th	14:00-15:40	B1-302	Group Meeting 2 (Write Report Section 1-4)
Monday 4nd	16:00-17:40	A310	Introduction to EXP2-1 Part2
Wednesday 6th	14:00-15:40	B1-302	Group Meeting 3 (Different sample comparison)
Thursday 7th	14:00-15:40	B1-302	Group Meeting 4 (Complete report Section 5-8)
Monday 11th	16:00-17:40	A310	Report submission