ANALYSIS OF ENGINEERING AND SCIENTIFIC DATA
(or How to deal with data)

Notices: Make sure you attend your tutorial at your allocated time and computer laboratory.

About This Lecture:
This lecture will
• calculate probabilities for normal distributions;
• introduce quantile-quantile plots and their use;
• discuss transformations of random variables and give rules for means and variances;
• introduce statistical inference; and
• introduce hypothesis testing and confidence intervals for the mean of a Normal distribution.

BEFORE THE LECTURE:
Read Lecture 3 notes.
Skim read sections 5.2.3, 5.3.1, 5.5.3, 5.5.5, 6.1.1, 6.1.2 and 6.2.1-6.2.3 of the text (don’t need to know everything in the text, but it can help).

CALCULATING NORMAL PROBABILITIES
Example:
The net weight of a jar of fruit is assumed to come from a normal distribution with mean 137.2g. and with standard deviation 1.6g. If the label weight is 135.0g., find the probability that a jar contains less than the label weight:

Let \( W \) = net weight of jar; then we have \( W \sim N(137.2, 1.6^2) \)

and so, \[
\Pr(W < 135) = \Pr\left(\frac{W - 137.2}{1.6} < \frac{135 - 137.2}{1.6}\right)
\]
\[= \Pr(Z < -1.375) = 0.08\]

So about 8% of jars are below the label weight.

If we want to find the probability that a jar contains within 1g. of the label weight, then we want:

\[
\Pr(134 < W < 136) = \Pr\left(\frac{134 - 137.2}{1.6} < \frac{W - 137.2}{1.6} < \frac{136 - 137.2}{1.6}\right)
\]
\[= \Pr(-0.75 < Z < -2.00) = 0.2266 - 0.0228 = 0.2038\]

Sketch:
Suppose the mean can be adjusted and it is decided to adjust it so that only 1% of jars have net weight below the label weight of 135.0g. We want \( \mu \) such that

\[
0.01 = \Pr(W < 135) = \Pr \left( \frac{W - \mu}{1.6} < \frac{135 - \mu}{1.6} \right) = \Pr \left( Z < \frac{135 - \mu}{1.6} \right)
\]

Thus, \( \frac{135 - \mu}{1.6} = -2.326 \) and so \( \mu = 138.7 \)g

A better alternative is to **reduce the variability** (\( \sigma \)).

**QUANTILE-QUANTILE PLOTS**

Histograms can be misleading as an indicator of shape.

Or: investigate whether the cumulative proportion of data values changes in the same way as the normal probabilities (areas under the pdf) do when going from the smallest data value to the largest data value.

Do this by plotting \( N(0,1) \) quantiles vs data quantiles:

**Example:**

10 breaking strengths of a paper towel were taken:

<table>
<thead>
<tr>
<th>Data</th>
<th>Prob</th>
<th>z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7583</td>
<td>0.05</td>
<td>-1.64485</td>
</tr>
<tr>
<td>8527</td>
<td>0.15</td>
<td>-1.03643</td>
</tr>
<tr>
<td>8572</td>
<td>0.25</td>
<td>-0.67449</td>
</tr>
<tr>
<td>8577</td>
<td>0.35</td>
<td>-0.38532</td>
</tr>
<tr>
<td>9011</td>
<td>0.45</td>
<td>-0.12566</td>
</tr>
<tr>
<td>9165</td>
<td>0.55</td>
<td>0.12566</td>
</tr>
<tr>
<td>9471</td>
<td>0.65</td>
<td>0.38532</td>
</tr>
<tr>
<td>9614</td>
<td>0.75</td>
<td>0.67449</td>
</tr>
<tr>
<td>9714</td>
<td>0.85</td>
<td>1.03643</td>
</tr>
<tr>
<td>10688</td>
<td>0.85</td>
<td>1.64485</td>
</tr>
</tbody>
</table>

If the relative distances between each consecutive pair of data values matches the relative distances between each consecutive pair of \( z \)-values, then the shape of the data matches the shape of a normal distribution.

So, if the shape does match then a plot of the \( z \)-values against (ordered) data values, gives (approximately) a straight line.

**MINITAB plots** probability values (not \( z \)-values) so giving a probability plot rather than a Q-Q plot.

Very small data set so hard to reject normality.
Examples

Histograms and normal plot (quantile-quantile plot for normality) for SHIFTS:

With 6 intervals (symmetric) and 9 intervals (skew).

Shift 2 (left skew)

The plot loses linearity at top - non-normal.

Diameters (bi-modal)

The plot loses linearity everywhere - non-normal.

NB: Obtain a straight line if data shape and theoretical shape "match" (ie rates of change match).

Example Very right skew data:

There is very little linearity – very non-normal.
TRANSFORMATIONS

By **transforming** data which does not seem to fit the normal distribution, it may be possible to obtain data which does look normal.

For example, by taking logs of data which is right-skewed, we may obtain data which looks normal (the untransformed data would then come from a log-normal distribution).

LINEAR TRANSFORMATIONS

Often, **linear combinations** of normal data are taken. The new variable will still have a normal distribution, but we need to know its mean and variance:

If \( a, b, a_1, a_2, \ldots, a_n \) are constants and recall that the mean of \( X \) is \( \mu = E(X) \), then

\[
\text{Mean: } E(aX + b) = aE(X) + b
\]

\[
E(a_1X_1 + a_2X_2 + \ldots + a_nX_n)
= a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)
\]

Recall that the variance of \( X \) is \( \sigma^2 = Var(X) \); then

\[
\text{Variance: } Var(aX + b) = a^2Var(X)
\]

\[
Var(a_1X_1 + a_2X_2 + \ldots + a_nX_n)
= a_1^2Var(X_1) + a_2^2Var(X_2) + \ldots + a_n^2Var(X_n)
\]

The last variance result is only true if there is no "association" between data values (data values are said to be independent of each other).

Example

Frequently, the units of a random variable may need to be changed (e.g. inches to cm, lbs to kilograms, Fahrenheit to Centigrade).

Example

Standardising normal random variables:

\[
E\left( \frac{X - \mu}{\sigma} \right) = \frac{E(X) - \mu}{\sigma} = 0
\]

\[
Var\left( \frac{X - \mu}{\sigma} \right) = \frac{Var(X)}{\sigma^2} = 1
\]

Example

Assume the width \( (X) \) of a steel plate is random, with mean 0.1489cm and variance \( 6.9 \times 10^{-7} \) cm\(^2\), and that the machined slot \( (Y) \) on a steel block is random (and unrelated to \( X \)) with mean 0.1546cm and variance \( 1.04 \times 10^{-6} \) cm\(^2\). (This is a tolerance analysis problem.)

If \( U = Y - X \) is the clearance, then \( U \) is random with mean and variance

\[
E(U) = E(-1 \times X + 1 \times Y) = -1 \times E(X) + 1 \times E(Y) = -0.1489 + 0.1546 = 0.0057
\]

\[
Var(U) = Var(-1 \times X + 1 \times Y) = (-1)^2 \times Var(X) + 1^2 \times Var(Y) = 6.9 \times 10^{-7} + 1.04 \times 10^{-6} = 1.73 \times 10^{-6}
\]

with standard deviation 0.0013cm.
Example

Find the mean and variance for a sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_{i} = \frac{1}{n} X_1 + \frac{1}{n} X_2 + \ldots + \frac{1}{n} X_n$$

In this case, $a_i = \frac{1}{n}$ for all $i$, so

$$E(\bar{X}) = \frac{1}{n} E(X_1) + \frac{1}{n} E(X_2) + \ldots + \frac{1}{n} E(X_n)$$

$$= \frac{1}{n} \mu + \frac{1}{n} \mu + \ldots + \frac{1}{n} \mu = \frac{n}{n} \mu = \mu$$

$$Var(\bar{X}) = \left(\frac{1}{n}\right)^2 Var(X_1) + \ldots + \left(\frac{1}{n}\right)^2 Var(X_n)$$

$$= \left(\frac{1}{n}\right)^2 \sigma^2 + \ldots + \left(\frac{1}{n}\right)^2 \sigma^2$$

$$= n\left(\frac{1}{n}\right)^2 \sigma^2 = \frac{n}{n} \sigma^2 = \frac{\sigma^2}{n}$$

Thus: $E(\bar{X}) = \mu$, $Var(\bar{X}) = \frac{\sigma^2}{n}$

If the $X_i$'s are normal, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Notes:

- The sample mean's variance ($\sigma^2/n$) decreases as the sample size gets larger; ie, the sample mean becomes more precise for larger samples.

- Even if the original distribution is not normal, the normal distribution may be used to obtain approximate probabilities for the sample mean provided the sample size is large enough (called the Central Limit Theorem).

STATISTICAL INFERENCE

AIM:

To quantify the reliability of data based conclusions (ie how precise are they).

Two main aspects:

- Parameter estimation - how precise are the estimates.

- Quantify the chance of errors in decision making.

CONFIDENCE INTERVALS

- A confidence interval for a parameter is a random interval (data-based) which contains the parameter with some (high) probability.

- This leads to reasonable confidence that the interval contains the true parameter value.

CONFIDENCE INTERVALS for the MEAN ($\mu$)

Case 1:

A sample, $X_1, X_2, \ldots, X_n$, from a normal distribution with known standard deviation ($\sigma$) is obtained and estimation of $\mu$ is required.

Use the sample mean, $\bar{X}$, to make inferences about $\mu$, using the fact that:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

We want an interval of the form

$$\bar{X} - a < \mu < \bar{X} + a$$

to contain $\mu$ with high probability, say 0.95.
That is, \( \Pr(\bar{X} - a < \mu < \bar{X} + a) = 0.95 \)

Re-organising this interval and standardising, we have:

\[
\Pr(\mu - a < \bar{X} < \mu + a) = 0.95
\]

and so

\[
\Pr\left( \frac{-a}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{a}{\sigma/\sqrt{n}} \right) = 0.95
\]

Since \( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \), we have:

\[
\frac{a}{\sigma/\sqrt{n}} = 1.96
\]

and so

\[
a = 1.96 \frac{\sigma}{\sqrt{n}}.
\]

Once we observe the sample mean and substitute it in, the interval is called a 95% confidence interval.

Note: If a different probability to 0.95 was used, the value 1.96 would change.

Example:

Suppose that 47 jars of fruit are chosen (at random) and the net weights measured. Assuming that \( \sigma = 1.6 \), obtain a 90% confidence interval for \( \mu \). Starting with the last statement above:

\[
\Pr\left( -1.645 < \frac{\bar{X} - \mu}{1.6/\sqrt{47}} < 1.645 \right) = 0.90
\]

then re-organising the inequalities gives:

\[
\Pr\left( \frac{\bar{X} - 1.645 \times 1.6}{\sqrt{47}} < \mu < \frac{\bar{X} + 1.645 \times 1.6}{\sqrt{47}} \right) = 0.90
\]

Interpreted this as “the interval \( \left( \bar{X} \pm 1.645 \times \frac{1.6}{\sqrt{47}} \right) \)
contains \( \mu \) with probability 0.9”

If the sample mean was observed to be 138.2g, then the 90% confidence interval for \( \mu \) would be:

\[
(138.2 - 0.3839, 138.2 + 0.3839) = (137.82, 138.58).
\]

We say this as:

- “there is 90% confidence that \( (137.82, 138.58) \)
  contains the true value of \( \mu \)”.

We interpret this interval as follows:

- “if we calculated a very large number of 90% confidence intervals, then we expect to have about 90% of them containing the true value of \( \mu \)”
  (though we don’t know which ones do and which ones don’t!).

Alternatively, suppose that it was desired that the sample mean, \( \bar{X} \), be within 0.3 grams of the true mean, \( \mu \), and we wish to know how often this will occur (a specification problem). We want to find:

\[
\Pr(-0.3 < \bar{X} - \mu < 0.3)
\]

\[
= \Pr\left( \frac{-0.3}{1.6/\sqrt{47}} < Z < \frac{-0.3}{1.6/\sqrt{47}} \right)
\]

\[
= \Pr(-1.28 < Z < 1.28)
\]

\[
= \Pr(Z < 1.28) - \Pr(Z < -1.28)
\]

\[
= 0.90 - 0.10 = 0.80
\]
This probability statement can be written in confidence interval form as

\[ \Pr(\bar{X} - 0.3 < \mu < \bar{X} + 0.3) = 0.80 \]

simply by re-arranging the two inequalities. So an 80% confidence interval would be:

\[ (138.2 - 0.3, 138.2 + 0.3) = (137.9, 138.5). \]

In general, a 100(1-a)% confidence interval for \( \mu \) would be:

\[ (\bar{X} - z_{\frac{a}{2}} \times \frac{\sigma}{\sqrt{n}}, \ \bar{X} + z_{\frac{a}{2}} \times \frac{\sigma}{\sqrt{n}}) \]

where

\[ \Pr(Z > z_{\frac{a}{2}}) = \frac{a}{2} \]

**Notes:**
- It is important to have some idea of just how accurate (precise) the estimate is.
- The precision (reliability) can be judged by the confidence interval's width (and the confidence coefficient) in relation to specifications and so on.
- For very large samples, \( \sigma \) in the above expressions can be replaced by the sample standard deviation, \( S \).
- Correct interpretation of a confidence interval:
  - If we have 100 95% confidence interval, then we “expect” about 95 of them to cover the true value.
  - BUT: which ones do is unknown!

**HYPOTHESIS TESTS**

An hypothesis test is a test of two possible hypotheses for a parameter (often "status quo" versus some alternative, but not always) on the basis of data.

The probabilities of choosing one hypothesis when the other is true are used to quantify the errors of choosing the wrong conclusion and, therefore, how well the test is doing.

**Uses:**
- To see if a “product” is within specifications.
- To compare two or more “methods” (processes, machines, products, raw materials etc).
- To determine if a process is out of control.
- To investigate different factors affecting response variables

**Example**

In a manufacturing process, it is thought that there should be no more than 5% of the items produced which are defective. A sample of 100 items is taken at random and the number of defective items counted.

Suppose that 8 items out of the 100 were defective. Is this evidence to suggest that the percentage of defective items being produced is larger than 5%?

**Make the following assumptions:**
- Each time an item is included in the sample, there is the same chance (probability, say \( p \)) it is defective.
- Knowing a given item is defective does not change the probabilities that any other item is defective.
Note: The second assumption is called independence and would be reasonable if we take a random sample.

Under these assumptions, and taking \( p \) to be 0.05 (what it is thought to be), the probabilities for different numbers of defectives (say, \( X \)) would be:

\[
\begin{array}{cccccc}
    x & 7 & 8 & 9 & 10 & 11 \\
    \Pr(X = x) & 0.1060 & 0.0649 & 0.0349 & 0.0167 & 0.0072 \\
\end{array}
\]

Notes:

• These are obtained from the Binomial theorem and so this distribution of probabilities is called a Binomial distribution.

• The random variable, \( X \), is a discrete random variable so we can have probabilities for individual values; we are using a different type of probability model here to the Normal distribution.

These are not the probabilities we want as we really want to know how likely we were to get a value which was too large for a \( p \) of 5%. So:

\[
\begin{array}{cccccc}
    x & 7 & 8 & 9 & 10 & 11 \\
    \Pr(X \geq x) & 0.3840 & 0.2340 & 0.1280 & 0.0631 & 0.0282 \\
\end{array}
\]

Obtaining a value as large as 8 is not that unlikely. But, if there had been 11 or more defectives, then this would provide evidence for saying the percentage of defectives being made is more than 5%.

How can this be formalised?

How do we determine that “unlikely” is unlikely enough?

TERMINOLOGY AND NOTATION:

We wish to answer the following question:

• Is there enough evidence to indicate that the percentage of items being produced exceeds 5%?

Specify hypotheses:

• \( H_0: p = 0.05 \) called the null hypothesis:
  
  We stick with this unless the data provides evidence against it in favour of the alternative.

• \( H_A: p > 0.05 \) called the alternative hypothesis:
  
  We only take this if the data supports it enough.

Determine test procedure:

• Reject \( H_0 \) in favour of \( H_A \) if \( X \) is too large for \( p = 0.05 \).

What does too large mean? Work this out in terms of how likely the observed number of defectives is when \( H_0 \) is true: (ie when \( p = 0.05 \)):

• Calculate probability of observing the obtained value of \( X \) or “worse” as regards \( H_0 \), to see if the data supports a change from \( H_0 \) to \( H_A \):

  Given the number of defectives is 8 and assuming that \( p = 0.05 \), we calculate \( \Pr(X \geq 8) = 0.2340 \)

• Is this probability unreasonably small? (If so, then the observed value of \( X \) would seem unlikely when \( H_0 \) is true, so probably \( H_A \) is true.)

  At 0.2340, this is not small so the data does not provide enough evidence against the null hypothesis is favour of the alternative; so stick \( p = 0.05 \).
LECTURE SUMMARY:

Write down seven steps in finding a probability or quantile for a (non-standard) normal distribution:

What are quantile-quantile plots and how are they used:

Give one way of dealing with non-normal data:

Write down the mean and variance for the following linear transformations of random variables:
(a) Simple linear transformation of one random variable:

(b) Difference of two random variable:

(c) Sample mean:

What does it mean to “standardise” a random variable?

What can the Central Limit Theorem be used for?

Write down two general objectives of statistical inference:

What are confidence intervals used for?

How should confidence intervals be interpreted?

What effect does the sample size have on the width of a confidence interval?
What effect does the standard deviation have on the width of a confidence interval?

What effect does the confidence co-efficient have on the width of a confidence interval?

What are hypothesis tests used for?

Explain the difference between the null hypothesis and the alternative hypothesis.

Explain the nature of the test procedure for testing hypotheses.

SAMPLE QUESTIONS FOR LECTURE 3

(1) In a water purification process, a machine adds a certain chemical every hour, with the amount added dependent on the average flow rate for the hour just prior to addition. Assume that the average flow rate over an hour is a normally distributed random variable, with a mean of 3.2 kilolitres and a standard deviation of 0.4 kilolitres. Suppose that the machine is stopped if this flow rate exceeds 4.1 kilolitres.

Let \( X \) denote the average flow rate for a given hour, so that \( X \sim N(3.2, 4.1^2) \).

(a) Find the probability that the average flow rate for this hour exceeds 4.1 kilolitres.

(b) Assume that the flow rate can be controlled so that its standard deviation can be changed. If \( \sigma \) is now the standard deviation for the average flow rate, find the value of \( \sigma \) so that there is a probability of 0.001 that the average flow rate for the hour exceeds 4.1 kilolitres.

(c) Suppose that 10 kilograms of chemical is added per kilolitre. It is desired to investigate how much chemical is required for a nine hour period (assuming the average flow rates for each hour are independent over this period - probably not a reasonable assumption!).

Let \( X_i \) denote the average flow rate for the \( i \)-th hour, so that \( X_i \sim N(3.2, 4.1^2) \).

Write down an expression for the total amount of chemical required in terms of the \( X_i \)'s.

Find the mean and standard deviation of the total amount of chemical required for this period.
(2) Suppose that conditions affecting the average flow rate over an hour have changed in Question (1). Although it is thought sensible to assume the standard deviation of 0.4 kilolitres still applies, it is expected that the mean flow rate has increased from the original mean of 3.2 kilolitres. To estimate what it might be now, a random sample of 9 flow rates is taken:

(a) Given that \( \sum_{i=1}^{9} X_i = 31.50 \), obtain a 95% confidence interval for the mean flow rate. Does it appear that the original flow rate of 3.2 is reasonable?

(b) Repeat (a) using a 99% confidence interval. Does it appear that the original flow rate of 3.2 is reasonable? Do you come to the same conclusion as in (a)?

(c) Repeat (a) assuming that the population standard deviation is now 0.6. Does it appear that the original flow rate of 3.2 is reasonable? Do you come to the same conclusion as in (a)?

(d) Suppose that it is now thought that the standard deviation has changed. Calculate the sample variance for this data set, given that \( \sum_{i=1}^{9} X_i^2 = 111.05 \).

(3) A manufacturing plant makes a certain component. Near the start of a batch run, a random sample of these components is obtained. Using accelerated testing, the time from when each is put in use till the first time it fails (the component’s lifetime) is obtained. It is desired to model the lifetime distribution in order to assess the reliability of the component. Near the end of the batch run, a second random sample is taken and the lifetimes obtained similarly.

The data is in a worksheet named stat3.mtw (on the Lecture Materials page). It contains the lifetimes for the first sample in a column (c1) called Lf1. The square roots of these lifetimes are in a column (c2) called sqrt(Lf1). The square roots of the lifetimes from BOTH samples are in a column (c3) called sqrt(Lf).

If you have access to MINITAB, use the instructions after the exercises to reproduce the output below.

(a) The histograms and normal plots for each data set are below. On each histogram, draw in the normal density curve (check where these should be after doing the MINITAB instructions). For each of the three data sets, describe the shape of the data from the histogram and discuss whether normality is appropriate from the normal plot (indicate the shape of the normal plot):
Square roots of lifetimes for both samples

Histogram of $\sqrt{L_f}$

(b) Discuss what may have caused the shape of the histogram of square roots of both samples (this is not a statistics question!).

(c) Discuss whether it is sensible to combine the two samples as has been done.

(4) Consider Question 1 in the Lecture 2 Exercises. If 7 was added to all the flow rates for stream 3, what effect would be seen in the mean, median, standard deviation and quartiles of the new values as compared to those for the original values.

If all the flow rates for stream 3 were multiplied by 2, what effect would be seen in the mean, median, standard deviation and quartiles of the new values as compared to those for the original values.

APPENDIX: Minitab Instructions for Question 3

Copy the worksheet `stat3.mtw` to an appropriate place and retrieve it.

Have a look in the worksheet window (now entitled `stat3.mtw`) to see the data. You can now proceed to investigate the shape of the data sets and whether normality is appropriate:

(a) Obtain histograms of the data in $L_f$, $\sqrt{L_f}$ and $\sqrt{L_f}$:
Click on Stat, then Basic Statistics, then Display Descriptive Statistics. Enter $L_f$, $\sqrt{L_f}$ and $\sqrt{L_f}$ under Variables. Click on Graphs and Histogram of data, with normal curve. Click ok. Click ok.

(b) Obtain normal plots for the data in $L_f$, $\sqrt{L_f}$ and $\sqrt{L_f}$:
Click on Stat, then Basic Statistics, then Normality Test. Enter $L_f$ for Variable. Click ok. Repeat the previous sequence but enter $\sqrt{L_f}$ for Variable. Click ok.
Repeat the previous sequence but enter $\sqrt{L_f}$ for Variable. Click ok.

Don't forget to exit Minitab and to logoff at the end of the session.