The Normal Distribution

Lecture 15

Week 6
27th August, 2007
Today ...

- Define the normal distribution
- Graph the distribution
- Calculate probabilities for
  - \( P(X = x) \)
  - \( P(X < q) \)
  - \( P(X > q) \)
  - \( P(q_1 < X < q_2) \)
- Discover what is special about \( \bar{X} \pm s \)
- Standard normal distribution, \( Z \)
What is a normal distribution?

- Often when data is collected from an experiment, if of sufficient size, then it has been observed that more data points lie close to the mean than further away.
- If the distribution is graphed it takes an approximate bell shape.
- This has been generalised to the data following a pattern or model called the normal distribution.
- The equation of the curve can be derived from

\[
\phi(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right)
\]

Again you will be pleased to note that you do not have to learn this formula.
\[
\phi_x (x; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left\{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right\}
\]

You can see that the variables other than the \(x\), are the mean, \(\mu\) and variance, \(\sigma^2\).

This means that a normal distribution is dependent on these and so defined by these terms.

We then say that a random variable \(X\) is distributed normally \((\mu, \sigma^2)\) or \(X \sim N(\mu, \sigma^2)\).

Sometimes this is defined by using \(\sigma\) and not \(\sigma^2\).
Figure 15.1 Comparison of a “data-based” and normal density curves from CO₂ data

The normal density curve was calculated using the mean and variance from the data.

It is quite a close approximation for the data based curve.

If the association is deemed satisfactory then calculations based on the normal curve make life a lot easier.
\[ \mu \text{ specifies the centre of the curve and } \sigma \text{ regulates the spread of the curve.} \]

- The normal distribution appears often in statistics because it reliably represents the random pattern in many, but not all, continuous situations.

- The imperfections that can be seen in Fig 15.1 are of a smaller order of magnitude than the random or unexplained components themselves and such discrepancies will have a negligible impact on the interpretation.
Even though this model fits many data sets, each will have its own $\mu$ and $\sigma$.

These are called *parameters* and the information about the density function is contained in them.

Hence it is called a *parametric density function*.

Explain the relationships to $\mu$ and $\sigma$. 
There are many parametric density functions, like Gamma and Uniform distributions, but we will concentrate on the normal distribution in this unit.

There are nonparametric density estimates, too. They are defined from the data set only.

The density curve fitted to the Milikan data set in Lecture 13 is such an example.
Remember that for a normal distribution, $X \sim N(\mu, \sigma^2)$.

We need to estimate these parameters from the data set by using:

\[
\hat{\mu} = \bar{X} \\
\hat{\sigma} = s
\]

We say that the rv $X$ is distributed normally around $\mu$ with a sd of $\sigma$.

Note that $\mu$ and $\sigma$ have a ‘hat’ above them. This signifies that $\mu$ hat is an estimate of $\mu$ given by $X$ bar etc.
Comparing the nonparametric curve with the normal parametric curve for the Milikan data we can see that in this instance there is very little difference (seems reliable enough) and so the data can be modelled by the normal distribution.

All that is required then from the data is $X$ and $s$ rather than an equation for a probability distribution curve.

Note the normal curve is generated from the equation here and isn't quite the standard one due to rounding.
Note:

- As you see more data sets in both statistics and your other disciplines, you will see that data almost never conforms “nicely” to convenient mathematical forms or models and so approximations are necessary.

- When we say data have a certain distribution, we do not mean that the data are distributed *exactly* that way but that the distribution is a reliable approximation to represent the random features of it and the discrepancies are minor compared with the broad pattern of randomness.
Calculating normal distribution probabilities

- We could use a distribution graph to calculate probabilities like we have done before but it is just as easy – if not easier – to use our friendly tool R.
- In Rcmdr use

\textbf{Distributions >Continuous distributions > Normal distribution >Normal probabilities}

You will need to enter a quantile variable (x), mean and a standard deviation value. Select lower tail. This will then give you $P(X<x)$.

Remember for continuous distributions $P(X=x) = 0$
Example 15.1

Example: Do it with Milikan data, mean=4.781, s=0.015 and find P(X<4.76)
Output is P=0.08

The commands to use in R are similar to those we have already met for discrete distributions and we fill in the relevant values.

- `pnorm(q= , mean= , sd= ,lower.tail= )`
- If we want P(X>x) then we use lower.tail=F or select the upper tail option in Rcmdr.

See R Murison’s notes page 122 for the Milikan example R commands.
What are the electron charges lying 1 standard deviation from the mean? How much of the data is expected to lie within this interval?

\( \bar{X} = 4.781 \) and \( s = 0.015 \) so  
\[ \bar{X} - s = 4.766 = q_l \]
and
\[ \bar{X} + s = 4.796 = q_u \]

\[ P(q_l) = 0.16 \]
\[ P(q_u) = 0.84 \]
\[ P(q_l < X < q_u) = 0.84 - 0.16 = 0.68 = 68\% \]

The voltage for this data is measured in microvolts.
We can do the same exercise to find the probability for finding data within 2 and 3 sd’s from the mean.

(X ± 2 s) = 4.781 ± 2 * 0.015
  = 4.781 ± 0.03
  = (4.751, 4.811)

(X ± 3 s) = 4.781 ± 3*0.015
  = 4.781 ± 0.045
  = (4.736, 4.826)
Using Rcmdr we get

\[ P(X \pm 2s) = P(4.751 < X < 4.811) \]
\[ = 0.97724987 - 0.02275013 \]
\[ = 0.9544997 \]
\[ \approx 95\% \]

\[ P(X \pm 3s) = P(4.736 < X < 4.826) \]
\[ = 0.998650102 - 0.001349898 \]
\[ = 0.9973002 \]
\[ \approx 99.7\% \]
Figure 15.5 More connections of $\mu$ with $\sigma$

$$P(X \pm 2 \sigma) = P(4.751 < X < 4.811) \approx 95\%$$

$$P(X \pm 3 \sigma) = P(4.736 < X < 4.826) \approx 99.7\%$$

Notice that there is 50% = $P(X < \mu)$ and 50% = $P(X > \mu)$

Mean = median for normal distributions.
The standard normal distribution

- If you look at Rcmdr you see that the default values for $\mu$ and $\sigma$ are 0 and 1 respectively.

- Prior to computing tools becoming available we had to rely on tables to calculate probabilities for given $q$ values or the reverse. We could find a $q$ from a given probability.

- To do this we needed to bring all normal distributions back to a standard so we didn’t need a table to cover every situation of $\mu$ and $\sigma$. 
To do this for any rv $X$, we need to subtract the mean and divide by the standard deviation.

That is \[ Z = \frac{X - \mu}{\sigma} \]

Recall that $E(aX) = aE(X)$
$E(a) = a$
$E(X + Y) = E(X) + E(Y)$

Hence if $E(X) = \mu$
then $E(Z) = \frac{E(X) - \mu}{\sigma}$
\[ = 0 \]
For the variance recall

\[ \text{Var}(aX) = a^2 \text{var}(X) \]
\[ \text{Var}(b) = 0 \]
\[ \text{Var}(X + c) = \text{var}(X) \]

So if \( \text{var}(X) = \sigma^2 \) then

\[ \text{Var}(Z) = \frac{1}{\sigma^2} \times \text{var}(X) + 0 = \frac{1}{\sigma^2} \times \sigma^2 \]

\[ = 1 \]

Expand these on the board.
Externals write the equations out in full to prove them to yourself.

You do not need to learn these proofs only the outcomes: \( E(Z)=0 \) and \( \text{Var}(Z) = 1 = \text{sd}(Z) \)
This result allows any normal data to be converted to standard normal data and the probabilities can be worked out from this special distribution.

They can also be compared because the units for this distribution are called standard deviation units!
From the Milikan data if \( X = 4.74 \) then

\[
Z = \frac{4.74 - 4.78}{0.015} = -2.67
\]

Using R (or Rcmdr)

\[
\text{pnorm(q=4.74,mean=4.78,sd=0.015)}
\]

0.0038

OR \[
\text{pnorm(q=-2.67,mean=0,sd=1)}
\]

0.0038
Figure 15.6 Normal distribution curve for Milikan data with standardised Z axis

Normal Distribution: $\mu = 4.781$, $\sigma = 0.015$

Note that the -2, -1, +1, +2 marks on the Z scale should be evenly spaced. I have tried to do this but not quite got it. Discrepancies due to rounding errors.
Using a standard deviation unit Z score to find X.

1st rearrange the formula then substitute in for the various symbols.