Statistics for Computing MA4413

Lecture 10

The Normal Distribution

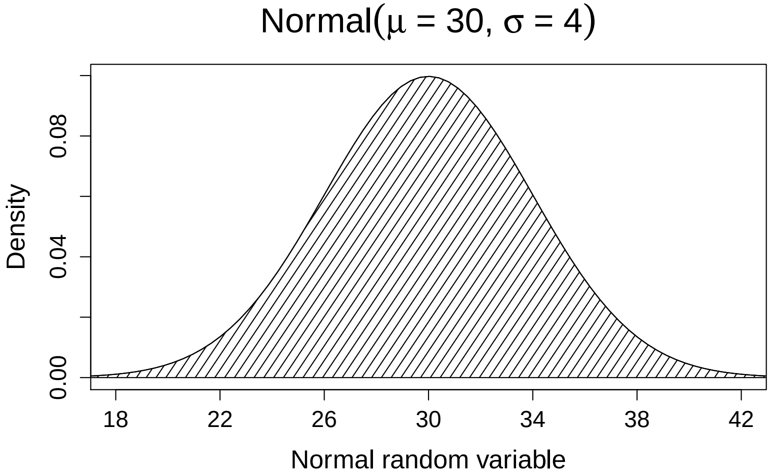
Kevin Burke

kevin.burke@ul.ie

Distributions Studied So Far

- Bernoulli(p)
  - An experiment where an event can occur with probability p.
  - $x \in \{0, 1\} \Rightarrow$ binary variable.
- Binomial(n, p)
  - The number of events in n Bernoulli trials.
  - $x \in \{0, 1, 2, \ldots, n\} \Rightarrow$ discrete variable.
- Poisson(λ)
  - The number of events in an interval of time / distance / space.
  - $x \in \{0, 1, 2, \ldots, \infty\} \Rightarrow$ discrete variable.
- Exponential(λ)
  - The time / distance / space between Poisson events occurring.
  - $t \in [0, \infty) \Rightarrow$ positive continuous variable.

Normal Distribution



A *continuous* distribution where the probability is distributed symmetrically around a central value, i.e., the mean.

Importance of the Normal Distribution

- Many biological, geographical, economic and demographic quantities are approximately normally distributed. Hence, it is a "normal" distribution, i.e., typical / common in practice.

- The mean of a *sample* of data is approximately normally distributed. This is known as the *central limit theorem* which underpins the most commonly used statistical testing procedures.

- Features of mass-produced products (e.g., dimensions, weight, volume etc.) are often normally distributed which is the basis of *quality control* procedures.

- Both the binomial (when n is large) and Poisson distributions (when λ is large) are approximately normal.

Normal Distribution

The **normal distribution** is used for *continuous* variables which are distributed symmetrically around the mean value.

$$\boxed{X \sim \text{Normal}(\mu, \sigma)}$$

$$\boxed{\Pr(X > x) = \int_x^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx}$$

where $\boxed{x \in (-\infty, \infty)}$

$$\boxed{E(X) = \mu}$$

$$\boxed{Var(X) = \sigma^2}$$

Normal Distribution

Clearly μ is the **mean** and σ is the **standard deviation** for a normally distributed random variable.

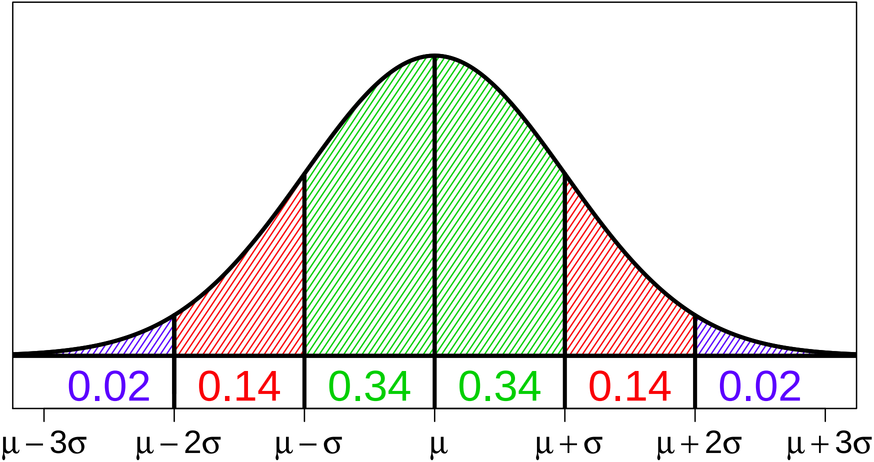
Although $x \in (-\infty, \infty)$ *in theory*, 99.7% of the probability is distributed to $x \in [\mu - 3\,\sigma, \mu + 3\,\sigma]$, i.e., $\Pr(\mu - 3\,\sigma < X < \mu + 3\,\sigma) = 0.997$.

Normal random variables are *continuous*. Thus, we calculate *greater than probabilities* (unlike the discrete cases). This is done via:

$$\Pr(X > x) = \int_x^{\infty} \frac{1}{\sigma\sqrt{2\,\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\, dx.$$

The above integral *cannot be done by hand*. We must use statistical tables or software.

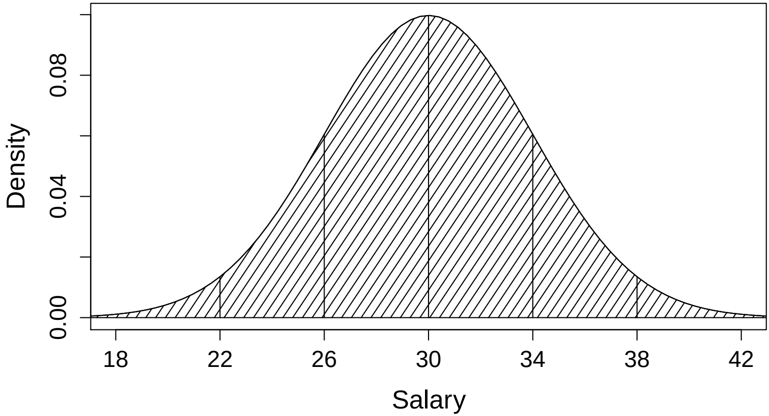
Approximating Normal Probabilities



- The above **approximate** probabilities can be used when we do not have stats tables.

Example: Salary

Let X be the salary (in thousands) for a particular type of job where $X \sim \text{Normal}(\mu = 30, \sigma = 4)$. Thus we know that:



Example: Salary

What is the probability that salary is greater than €26k?

$$\Pr(X > 26) \approx 0.34 + 0.34 + 0.14 + 0.02 = 0.84.$$

What is the probability that salary is between €26k and €34k?

$$\Pr(26 < X < 34) \approx 0.34 + 0.34 = 0.68.$$

What is the probability that salary is greater than €36k?

$$\Pr(X > 36) \approx \frac{0.14}{2} + 0.02 = 0.09.$$

(since €36k is halfway between €34k and €38k)

Normal Tables

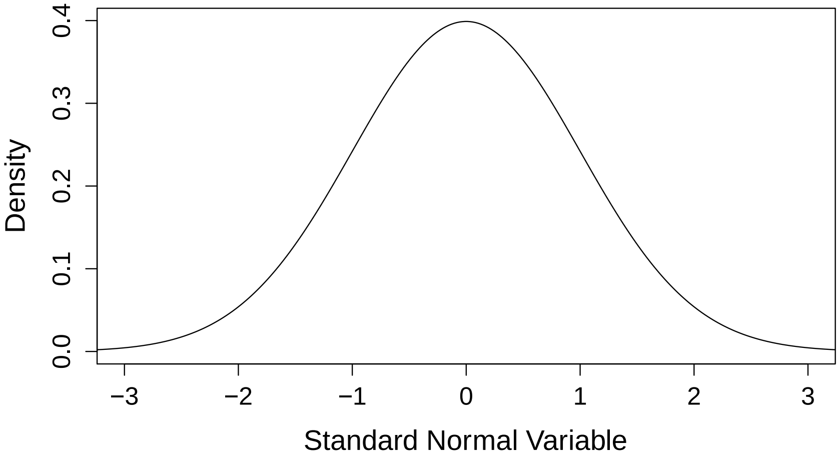
Using the “34-14-2” approximation is *not satisfactory*.

We can calculate normal probabilities *exactly* using the **normal tables**.

The tables show **greater than** probabilities corresponding to the **standard normal distribution**: Normal($\mu = 0, \sigma = 1$).

Having only the Normal($\mu = 0, \sigma = 1$) case tabulated is *not a limitation* since we can *standardise* any normal variable.

Standard Normal Distribution



- The letter Z is used to denote standard normal variables: $Z \sim \text{Normal}(\mu = 0, \sigma = 1)$

Normal Tables

The normal tables show **greater than** probabilities:

$$\boxed{\Pr(Z > z)}$$

but only for positive values of z.

We look up the z value in the row/column headings and to find the relevant probability.

Rows show the *first decimal place* of z and the **columns** show the *second decimal place*.

Normal Tables Examples

- $\Pr(Z > 0.40) = 0.3446$
- $\Pr(Z > 1.08) = 0.1401$
- $\Pr(Z > 2.00) = 0.02275$
- $\Pr(Z > 0.45) = 0.3264$
- $\Pr(Z > 1.80) = 0.0359$
- $\Pr(Z > 2.63) = 0.00427$

We calculate *less than* probabilities using the *complement* rule:

- $\Pr(Z < 0.40) = 1 - \Pr(Z > 0.40) = 1 - 0.3446 = 0.6554$
- $\Pr(Z < 1.08) = 1 - \Pr(Z > 1.08) = 1 - 0.1401 = 0.8599$

$\ldots$ etc.

Symmetry Rule

So we can calculate probabilities for positive z values, but what about **negative** values?

We must apply the **symmetry rule** for *standard normal variables*:

$$\boxed{\Pr(Z < -z) = \Pr(Z > z)}$$

or, similarly,

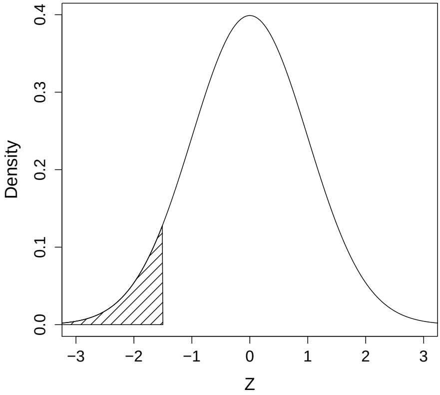
$$\boxed{\Pr(Z > -z) = \Pr(Z < z)}.$$

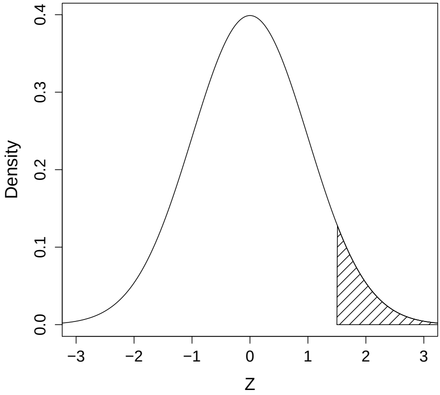
$\Rightarrow$ **Flip the inequality** symbol and **change the sign** of the z value.

(Note: this is *not* a general rule of probability - it can only be used for the standard normal distribution)

Symmetry Rule: Example 1

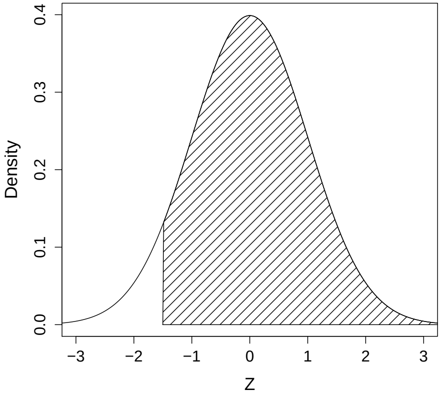
$$\Pr(Z < -1.5) \quad = \quad \Pr(Z > 1.5)$$

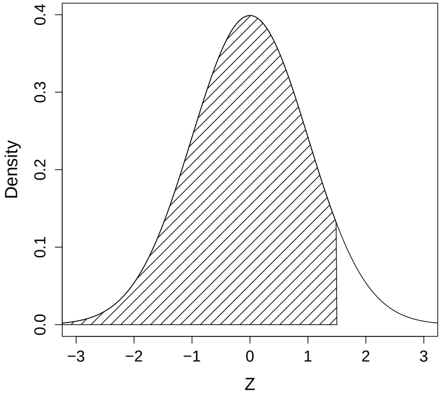




Symmetry Rule: Example 2

$$\Pr(Z > -1.5) \quad = \quad \Pr(Z < 1.5)$$





Normal Tables Examples

$$\begin{aligned} \Pr(Z < -1.74) &= \Pr(Z > 1.74) && \text{(symmetry rule)} \\ &= 0.0409. && \text{(using tables)} \end{aligned}$$

$$\begin{aligned} \Pr(Z > -0.60) &= \Pr(Z < 0.60) && \text{(symmetry rule)} \\ &= 1 - \Pr(Z > 0.60) && \text{(complement rule)} \\ &= 1 - 0.2743 = 0.7257. && \text{(using tables)} \end{aligned}$$

Normal Tables Examples

$$
\begin{aligned}
\Pr(-1.00 < Z < 0.85) &= \Pr(Z > -1.00) - \Pr(Z > 0.85) \\
&= \Pr(Z < 1.00) - \Pr(Z > 0.85) \\
&= [1 - \Pr(Z > 1.00)] - \Pr(Z > 0.85) \\
&= (1 - 0.1587) - 0.1977 \\
&= 0.8403 - 0.1977 \\
&= 0.6426.
\end{aligned}
$$

Question 1

Calculate the following:

a) $\Pr(Z > 0.83)$.

b) $\Pr(Z < 1.05)$.

c) $\Pr(1 < Z < 2)$.

d) $\Pr(Z < -1.8)$.

e) $\Pr(-1 < Z < 1)$.

f) The value of z such that $\Pr(Z > z) = 0.1$.

Standardising Normal Variables

For a normally distributed variable $X \sim \text{Normal}(\mu, \sigma)$, we can convert to a *standard normal variable* via:

$$\boxed{Z = \frac{X - \mu}{\sigma}} \sim \text{Normal}(\mu = 0, \sigma = 1),$$

i.e., we subtract the mean and divide by the standard deviation.

This process is called **standardising** and the resulting Z value is typically referred to as a ***Z* score**.

The Z score is the *number of standard deviations from the mean*.

Example: Salary

Earlier we had that $X \sim \text{Normal}(\mu = 30, \sigma = 4)$ and approximated probabilities (this is unsatisfactory).

Now we can standardise the variable:

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 30}{4}$$

and then use the normal tables to find the *exact* probabilities.

Example: Salary

What is the probability that salary is greater than €26k?

$$\text{standardise: } Z = \frac{26 - 30}{4} = \frac{-4}{4} = -1$$

$$\begin{aligned}
\Rightarrow \Pr(X > 26) &= \Pr(Z > -1) \\
&= \Pr(Z < 1) \\
&= 1 - \Pr(Z > 1) \\
&= 1 - 0.1587 \\
&= 0.8413.
\end{aligned}$$

Note that €26k is 1 standard deviation *below* the mean $\Rightarrow Z = -1$.

Example: Salary

What is the probability that salary is between €26k and €34k?

$$\begin{aligned}\Pr(26 < X < 34) &= \Pr(X > 26) - \Pr(X > 34) \\ &= 0.8413 - \Pr(Z > \tfrac{34-30}{4}) \\ &= 0.8413 - \Pr(Z > 1) \\ &= 0.8413 - 0.1587 \\ &= 0.6826.\end{aligned}$$

What is the probability that salary is greater than €36k?

$$\begin{aligned}\Pr(X > 36) &= \Pr(Z > \tfrac{36-30}{4}) \\ &= \Pr(Z > 1.5) \\ &= 0.0668.\end{aligned}$$

Question 2

Assume that 12V batteries are produced in a factory. Due to slight variations, the actual voltage is $X \sim \text{Normal}(\mu = 12, \sigma = 0.1)$, i.e., not every battery is exactly 12V. Calculate the following:

a) The proportion with more than 12.15V.

b) The proportion with less than 12.38V.

c) The proportion within the specification limits 12V $\pm$ 0.15V.

d) The value of x such that $\Pr(X < x) = 0.9$.

e) The value of x such that $\Pr(X < x) = 0.1$.

R Code

For the normal distribution we calculate *greater than* probabilities, i.e., $\Pr(X > x)$.

Examples:

`pnorm(26,mean=30,sd=4,lower=F)` gives `0.8413447`.

`pnorm(34,mean=30,sd=4,lower=F)` gives `0.1586553`.

`pnorm(36,mean=30,sd=4,lower=F)` gives `0.0668072`.

Compare this with slide 23.

R Code

We can *generate* normal random variables as follows:

Example:

`rnorm(100,mean=30,sd=4)`
generates 100 Normal($\mu = 30, \sigma = 4$) variables.

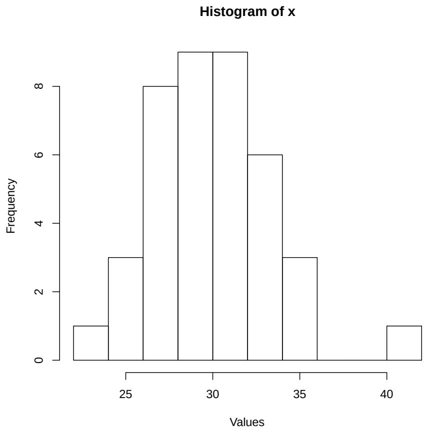
Checking Normality

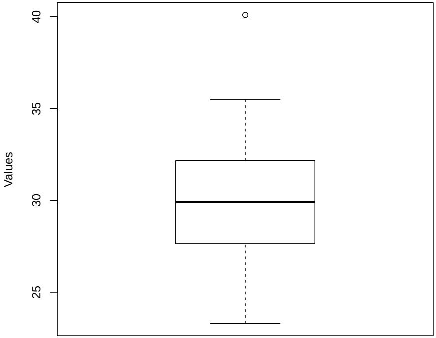
For a given set of data, it is often useful to check if the distribution looks approximately normal.

If this does turn out to be the case, we can calculate probabilities as shown on the previous slides.

We can also apply the *t test* to small samples that are approximately normal (more on this later).

Histogram / Boxplot





Together, the histogram and boxplot can tell us about the distribution of values. We see that the above looks approximately normal.

Q-Q Plot

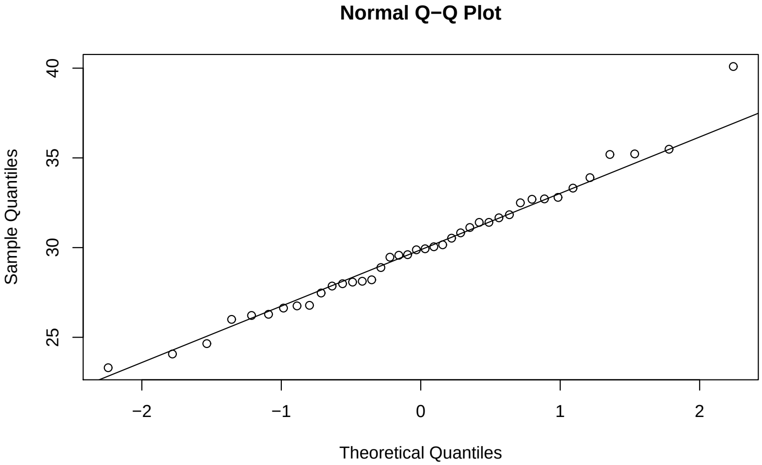
A more useful check for normality is the **quantile-quantile plot**.

The Q-Q plot compares the *quantiles* of the sample of data to those of a theoretical normal distribution.

If the data is approximately normally distributed, the quantiles match and the Q-Q points lie on a straight line.

Note: a quantile is a more general concept than *quartiles* which we studied earlier. In fact quartiles are the 4-quantiles.
(For your information: the 100-quantiles are known as percentiles)

Q-Q Plot



- The data appears to be approximately normally distributed apart from one outlier (compare with the histogram and boxplot on slide 28).

R Code

The graphs on the previous slides can be produced via:

```
set.seed(112187721)
x = round(rnorm(40, mean=30, sd=4),3)
hist(x, xlab="Values")
boxplot(x, ylab="Values")
qqnorm(x); qqline(x)
```

A sample of 40 Normal($\mu = 30, \sigma = 4$) variables were generated - try with different sample sizes.

Note the use of `set.seed` so that you can reproduce the exact data used in these slides - try the above without `set.seed`.

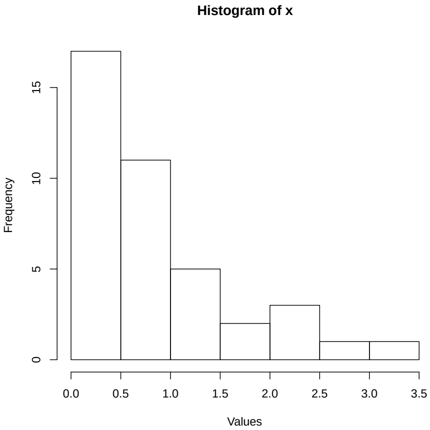
R Code

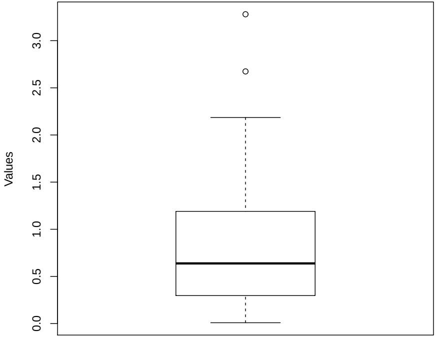
See what happens when we generate from an exponential distribution:

```
set.seed(112187721)
x = round(rexp(40, rate=1),3)
hist(x, xlab="Values")
boxplot(x, ylab="Values")
qqnorm(x); qqline(x)
```

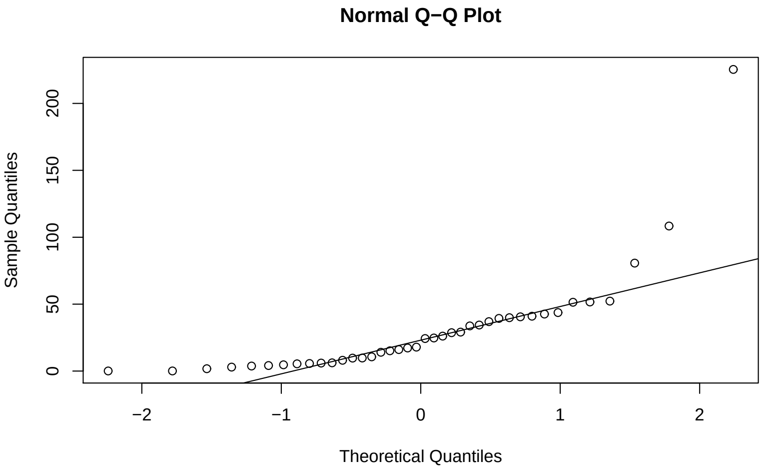
The output of the above code is shown on the next two slides. We can see that this data is not normally distributed.

Histogram / Boxplot





Q-Q Plot


Normal Q−Q Plot
Sample Quantiles
0
50
100
150
200
−2
−1
0
1
2
Theoretical Quantiles