Lecture 6

The Normal Distribution & The Inverse Problem and Normal Sampling Distributions

By
Dr Sean Maudsley-Barton and Abdul Ali

Manchester Metropolitan University

Aims

Manchester Metropolitan University

- Understand the concept of a continuous random variable and its probability distribution.
- Recognise the features of the **Normal distribution**.
- Use tables to find probabilities for the **standard Normal distribution**.
- Recognise that the **sample mean** and **sample totals** are also random variables with a probability distribution.
- Use Normal tables to solve problems involving the distribution of the sample mean.
- Appreciate the concept of the **central limit theorem**.
- Formulate and solve problems using the **inverse of the Normal CDF function**.

The Normal Distribution (Chapter 9)

Manchester Metropolitan University

Continuous Random Variables

- A continuous random variable is one that, in principle, can **take on any value** in a **given range or interval**.
- Of course, in practice, the **actual values** we observe are **always constrained by the accuracy** with which we can measure the variable.

Example 9.1 (*Continuous Random Variables*)

Examples of such variables might include,

- the time spent waiting for a bus
- the weight of the contents of a bag of sugar
- the height of an individual

The Normal Distribution

Manchester Metropolitan University

Examples of Continuous Distributions

- The **Normal distribution** is an example of a **continuous random variable**.

Density function:

- We describe the **probability distribution** for a continuous random variable in terms of a **mathematical formula**, known as a **density function**. This is the continuous analogue (i.e. *equivalent*) of the **discrete mass function**..
- We often **represent it graphically** to understand **how the probability is *distributed*** over the **values** of the variable.

Examples of Continuous Distributions

Example 9.1 (*Continuous Random Variables*)

If we imagine the density, $f(x)$ for a random variable X, say, being **represented by its graph**, the following **properties** hold,

1. The graph **cannot be negative**, i.e. it always **lies on or above the x-axis**.

2. The **area under the whole graph is 1**

3. The area under the graph between any two points a and b, is $P(a \leq X \leq b)$, i.e.

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Area between **a** and **b**

As we will see, **these properties** are very similar to those which hold for discrete random variables.

Examples of Continuous Distributions

Manchester Metropolitan University

Example 9.2 (*The Uniform distribution*)

The simplest **continuous distribution** is called the ***uniform distribution*** on **[0,1]**.

A uniform random variable, $\boldsymbol{X}$ say, has distribution,

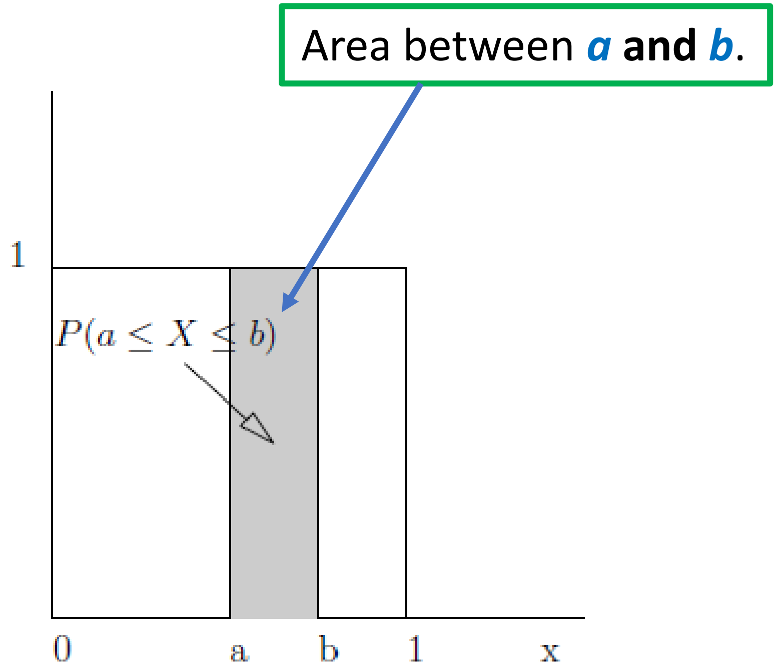
$$\boldsymbol{f(x)} = \begin{cases} \mathbf{1} & \mathbf{0 \leq \boldsymbol{x} \leq 1} \\ \mathbf{0} & \boldsymbol{otherwise} \end{cases}$$

- The **uniform distribution** is **equally likely to take on any value** in the **range [0,1].**
- Most calculators are able to produce such random variables (to 3 d.p.) - look for a button on your calculator labelled **Ran#** (often a second function).
- Such **uniform random numbers** have many **applications**, for example **in simulations**.

Examples of Continuous Distributions

Example 9.2 (*The Uniform distribution*)

The **graph** of this distribution looks like,



It should be clear from the diagram that,

1. The distribution is **never negative.**

2. The **area under the whole graph is 1** - it's a **square of side 1**.

3. Since all values are equally likely, the probability that the variable lies between any two values is just given by the area under the graph enclosed by those two values.

Examples of Continuous Distributions

Example 9.2 (*The Uniform distribution*)

- Clearly, the uniform distribution is a very simple kind of **continuous distribution** although it can be generalised.

- For example, we can **define a uniform distribution** on an **arbitrary range, [a, b],** say. Can you work out what the graph needs to look like in order to satisfy the requirements of being a distribution?

However, the **important concept**, that of the **representation of probability** for a **continuous distribution** as being an **area under the distribution graph**, is **common to all continuous distributions**.

The Normal Distribution

- The **Normal distribution** is the **most important** distribution in statistics.
- **Many variables** that are observed **will follow**, at least approximately, a **Normal distribution**.
- Moreover, it can be shown that, **under mild conditions**, whenever we **add together random variables, their distribution** will **tend towards** that of a **Normal variable**.

A random variable, $\mathbf{X}$ say, with a **Normal distribution** has **density function**, $\boxed{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu^2)}}$

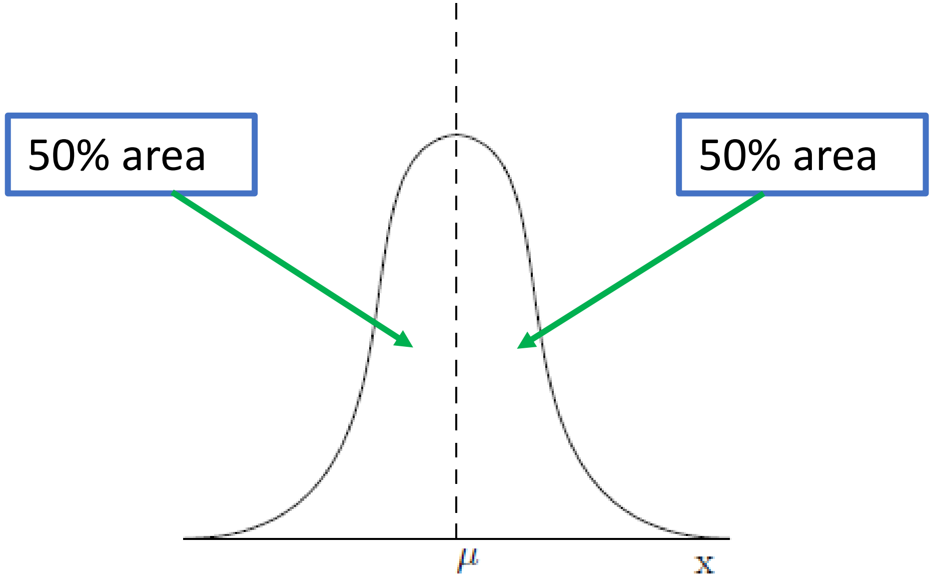
$$\boldsymbol{f(x) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) & -\infty < x, \mu < \infty, \sigma^2 > 0 \\ 0 & otherwise \end{cases}}$$

The **mean** (average) of this distribution is $\boldsymbol{\mu}$ and the **variance is** $\boldsymbol{\sigma^2}$.

However, in practice, we tend to make **more use of the standard deviation** $\boldsymbol{\sigma}$ $\left(\sqrt{\sigma^2} = \sigma\right)$.

The Normal Distribution

The probability distribution of a typical Normal distribution is shown in the diagram below.



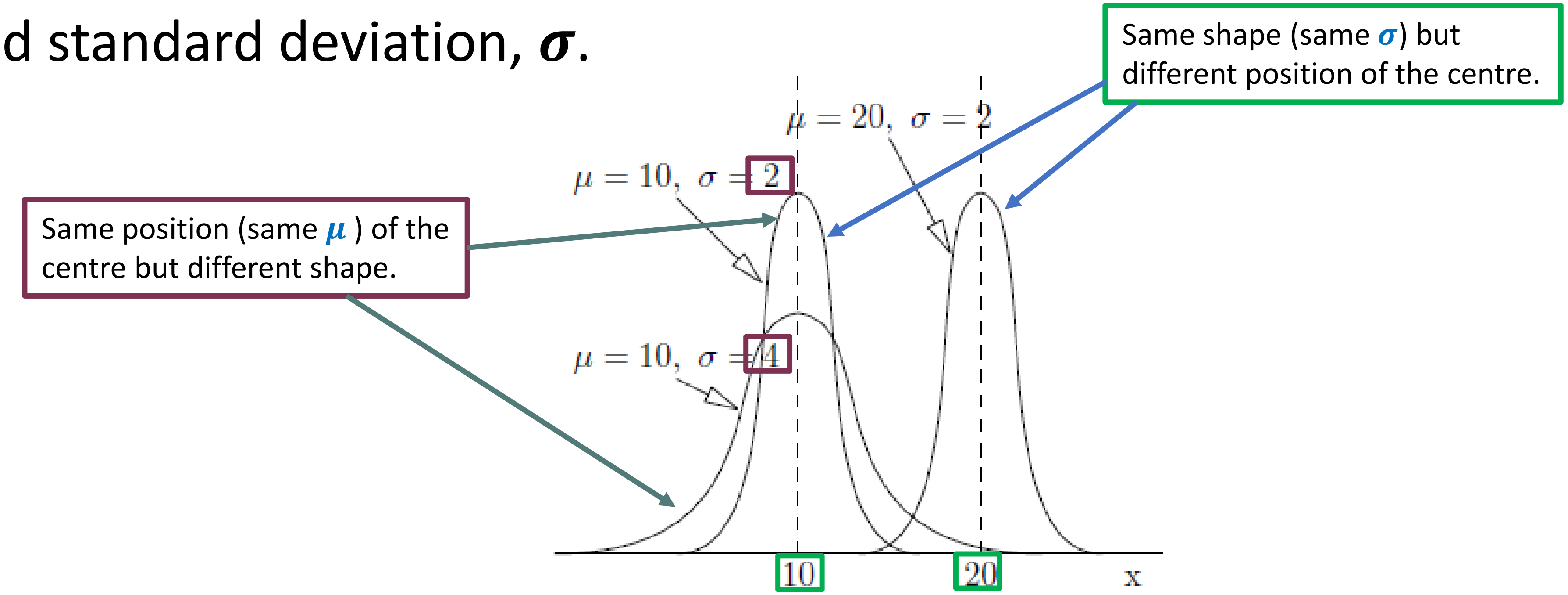
As you can see,

- It is a **bell-shaped** curve which is **symmetric about a value μ** ('mu').
- **Most of the probability**, i.e. the area under the curve, **is located around the mid-point μ** with only a relatively **small amount** at values **a long way from μ**. This suggests that we can expect **most observations to be close to the value μ** with only **a small proportion a long way from μ**.
- The role of σ is to **determine the spread or variability** of the random variable.

The interpretation of the **role of the mean and variance** in terms of random variables essentially mirrors that when dealing with samples of data.

The Normal Distribution

The following diagram shows **three Normal distributions** with different values of the mean, $\boldsymbol{\mu}$ and standard deviation, $\boldsymbol{\sigma}$.



- Changing the **value of μ** by itself simply **changes the position of the centre** of the curve.
- Changing the value of the **standard deviation (σ)** determines **how spread out the curve is** and, therefore, how variable the values of the random variable will be.

The Normal Distribution

As the **Normal distribution is determined by its mean and standard deviation**, we can denote the distribution in shorthand as follows.

- If a random variable, X say, follows a Normal distribution with mean μ and standard deviation σ, we write, $\boldsymbol{X \sim N(\mu, \sigma^2)}$.

Note that we **always refer to the variance** when specifying the Normal distribution, i.e. variance, $\boldsymbol{\sigma^2}$ **= s.d.**^{2} .

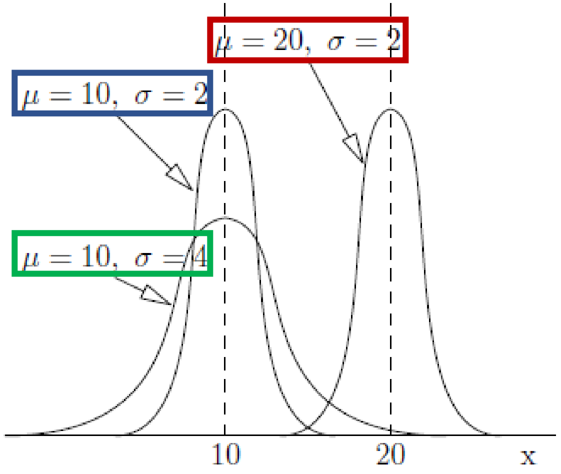
- Thus the **three distribution** in the diagram could be referred as,

$$\boldsymbol{N(10, 4), N(10, 16) \text{ and } N(20, 4)}$$

or as,

$$\boldsymbol{N(10, 2^2), N(10, 4^2) \text{ and } N(20, 2^2)}$$

Can you identify which is which?



The Standard Normal Distribution

Manchester Metropolitan University

The Normal distribution with μ = 0 and σ = 1 is referred to as the standard Normal distribution, and is usually denoted by the letter Z.

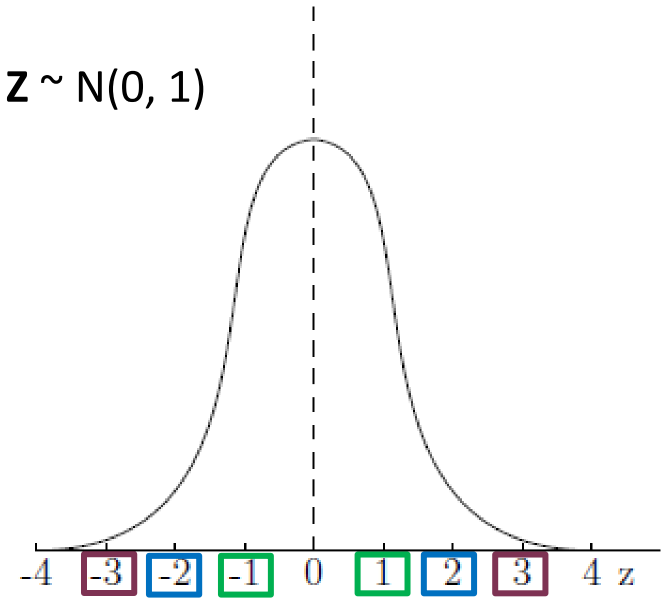
Clearly, the **density function** of the standard Normal is a **special case** of the Normal density shown in the figure above and is given by,

$$\boxed{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}$$

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) & -\infty < x < \infty \\ 0 & otherwise \end{cases}$$

The Standard Normal Distribution

A **graph of the standard Normal distribution** is shown below,



Area between,
$-1 < Z < 1 = 68.3\%$
$-2 < Z < 2 = 95.4\%$
$-3 < Z < 3 = 99.7\%$

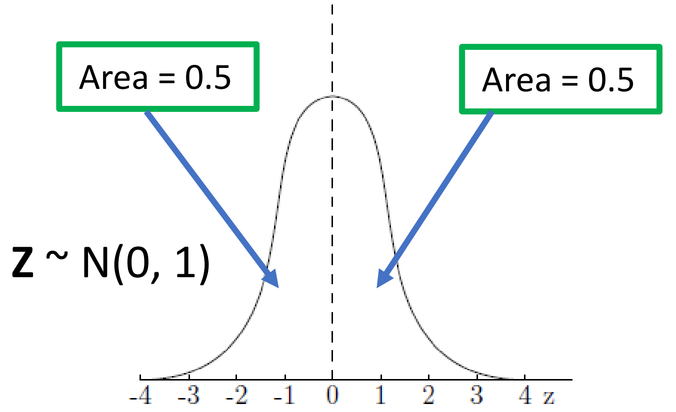
Notice that most of the area (99.7%) under the graph is **located between -3 and +3**.

The standard Normal is the distribution which is provided in the tables.

In the next few slides we will see how a **problem involving any Normal distribution can be expressed in terms of one involving the standard Normal.**

Using Standard Normal Tables

The **MMU tables give areas**, **i.e. probabilities**, **in the upper half of the distribution for z > 0**.



However, we can **find any area/probability** by observing:

1. The area under the whole graph is **1**.

2. The **graph is symmetric** about the **mid-point (μ)**.

This implies that the areas above and below the mid-point are both **0.5**.

3. The **law of complements** applies, e.g. $P(Z \geq z) = 1 - P(Z \leq z)$

However, it is generally useful to **draw a simple sketch diagram** of the distribution which shows the values required for a particular problem and the areas representing the probabilities.

Using Standard Normal Tables

Table Rules:

1. $P(Z > +z)$ - Found directly from table {**Note:** P(Z>a) = $P(Z \geq a)$}

2. $P(Z < +z) = 1 - P(Z > +z)$ - Complement law

3. $P(z_1 < Z < z_2) = P(Z \geq z_1) - P(Z \geq z_2)$ - Between two points a and b.

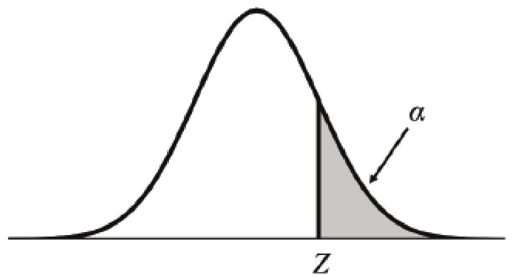
4. $P(Z < -z) = P(Z > +z)$ - Symmetry

5. $P(Z > -z) = 1 - P(Z > +z)$

Using Standard N

2.3 Normal Distribution

Tabulated values of $P(Z > z)$ where $Z = \dfrac{x - \mu}{\sigma}$.



Example 9.3 (*Standard No*

We will use tables to find the ... ndom variable, Z.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048

a) $P(Z \geq 2.45)$

i. Formulate lookup

$= P(Z > 2.45)$

$= P(Z > 2.4 + 0.05)$

ii. Look up the value

= **0.0071**

table

Complement law

$Z \geq b)$ - Between two points a and b.

metry

Applications of the Normal Distribution

Important points:

- **Normal distribution** was **introduced** in the slides above and a **special form** of the distribution was considered - the **standard Normal with mean 0 and variance 1**.

- In general, a **Normal distribution can be defined with any value** for the **mean** and any **positive value for the variance or standard deviation**.

- Given this, it is clearly **not possible to produce tables** for **each and every Normal distribution**.

- However, the **following theorem** shows that **we only ever need to have tables** of the **standard Normal** in order **to find any Normal probabilities**.

Using Standard Normal Tables

Manchester Metropolitan University

Theorem 9.1 (*Standardising Normal Variables*)

Suppose a random variable, $\boldsymbol{X \sim N(\mu, \sigma^2)}$, i.e. the **random variable $\boldsymbol{X}$** follows a Normal distribution with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ (its variance is σ^2). Then the **standardised variable $\boldsymbol{Z}$**,

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

and, in particular,

$$\boxed{\boldsymbol{P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)}}$$

for any values of $\boldsymbol{a}$ and $\boldsymbol{b}$.

This formula has a very simple **geometric interpretation**.

Recall that we can **represent probabilities** for continuous random variables **as the area**, between specified limits, under the distribution curve.

This **theorem** just says that the **area under any Normal curve** between limits $\boldsymbol{a}$ and $\boldsymbol{b}$ is always the **same as the area under the standard Normal curve** between the transformed limits $\boxed{\boldsymbol{\frac{a-\mu}{\sigma}}}$ and $\boxed{\boldsymbol{\frac{b-\mu}{\sigma}}}$.

Using Standard Normal Tables

Procedure:

The following **procedure** will help in solving problems, but you will find it becomes easier with practice.

1. Identify the **value of the mean and standard deviation** of your Normal distribution.

2. **Draw a sketch graph** of your distribution, indicating on it the **position of the *mean***. Mark approximately the values of **any limits** in your problem.

3. Transform the original limits by subtracting the mean and then dividing by the standard deviation (*NOT the variance*).

4. **Draw a sketch of the standard Normal distribution** with **mean marked at zero**.

5. The **shape of the area** enclosed by the **original and transformed** limits should look the same. If they don't you've made an error.

With these **procedures and properties**, and a little lateral thinking, **you can find probabilities for any Normal distribution** just from tables of the standard Normal.

Using Standard Normal Tables

Example 9.4 (*Apples*)

Suppose that the weight of a particular grade of apples is Normally distributed with **mean 100g** and **standard deviation 8g**. Let X denote the weight of a randomly selected apple, i.e. $\mathbf{X} \sim \boldsymbol{N}\ (\mathbf{100, 8^2})$, find

1. $P(X > 115)$
2. $P(X < 80)$
3. $P(105 < X < 112)$
4. $P(95 < X < 112)$

Firstly make sure you correctly identify the values μ = 100 and σ = 8. The steps involved are,

1.

$$P(X > 115) = P\left(Z > \frac{115 - 100}{8}\right)$$
$$= P(Z > 1.88)$$

$$\boxed{\boldsymbol{Z} = \frac{\boldsymbol{X} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{1})}$$

This quantity can be found directly from tables, i.e. $P(Z > 1.88) = \mathbf{0.0301}$.

Using Standard Normal Tables

Manchester Metropolitan University

Example 9.4 (*Apples*)

2.

$$P(X < 80) = P\left(Z < \frac{80 - 100}{8}\right)$$

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

$$= P(Z < -2.5)$$

$$= P(Z > 2.5) \qquad \text{(by symmetry)}$$

From tables, i.e. $P(Z > 2.5) = \mathbf{0.0062}$.

3.

$$P(105 < X < 112) = P\left(\frac{105 - 100}{8} < Z < \frac{112 - 100}{8}\right)$$

$$= P(0.63 < Z < 1.5)$$

$$= P(Z > 0.63) \ - \ P(Z > 1.5) \qquad \text{(difference of two) sets}$$

$= 0.2643 - 0.0668 =$ **0.1975**

Using Standard Normal Tables

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

Manchester Metropolitan University

Example 9.4 (*Apples*)

4. $$P(95 < X < 112) = P\left(\frac{95-100}{8} < Z < \frac{112-100}{8}\right)$$
$$= P(-0.63 < Z < 1.5)$$

But,

$P(Z < -0.63) + P(-0.63 < Z < 1.5) + \mathrm{P}(\mathrm{Z} > 1.5) = 1$ - the three areas comprise the **whole distribution**.

Alternate method: Using table rule 3 & 5.

$$P(-0.63 < Z < 1.5) = P(Z > -0.63) - P(Z > 1.5) \quad \{\text{Rule 3}\}$$
$$= 1 - P(Z > 0.63) - P(Z > 1.5) \ \{\text{Rule 5}\}$$
$$= 1 - 0.2643 - 0.0668 = \mathbf{0.6689}$$

We also have,

$P(Z < -0.63) = P(Z > 0.63)$ (by symmetry)

So that

$$P(-0.63 < Z < 1.5) = 1 - P(Z > 0.63) - P(Z > 1.5)$$
$$= 1 - 0.2643 - 0.0668 = \mathbf{0.6689}$$

Table Rules:

1. $P(Z > +a)$ - Found directly from table
2. $P(Z < +a) = 1 - P(Z > +a)$ - Complement law
3. $P(a < Z < b) = P(Z \geq a) - P(Z \geq b)$ - Between two points a and b.
4. $P(Z < -a) = P(Z < +a)$ - Symmetry
5. $P(Z > -a) = 1 - P(Z > +a)$

The Inverse Problem

In *Applications of the Normal Distribution* topic we saw how to solve problems which involved finding the probability that a Normally distributed random variable lay in a certain range.

The solution to this problem consisted of two steps:

1. standardising the value of the **original variable** to get a standard Normal variable.

2. using **tables of the standard Normal** distribution to find the required probability, recognising that the process of standardisation preserves areas, i.e. probabilities.

We may think of the process as follows:

$$\text{Original Value, } X \xrightarrow{\frac{x-\mu}{\sigma}} Z \longrightarrow \text{Probability ?}$$

The Inverse Problem

The inverse problem, as the name suggests, is simply the **same process but applied backwards**. We **start out with a probability** and seek to **find the value of the random variable** corresponding to that probability. Thus, since

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

$$\Rightarrow \quad \sigma Z = X - \mu$$

$$\Rightarrow \quad \boxed{\boldsymbol{X = \mu + \sigma Z}}$$

the inverse problem can be thought of as working through the following process,

Original Value, $X \xleftarrow{\boldsymbol{\mu + \sigma Z}} Z \longleftarrow$ Probability

As before a simple sketch graph of the problem is invaluable.

The Inverse Problem

Manchester Metropolitan University

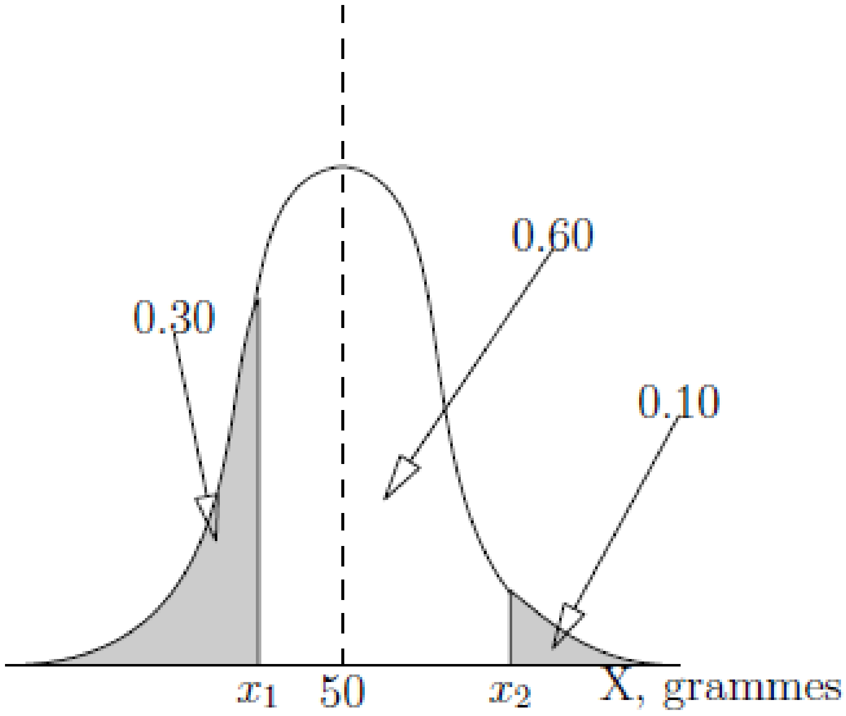
Example 10.3 (*Eggs*)

The weights of eggs laid by a particular breed of hens are Normally distributed with **mean 50g** and **standard deviation 5g**. An egg producer wants to classify eggs so that the **heaviest 10%** are classified as large and the **lightest 30%** classified as small. **The remaining 60%** are classified as medium. What weights should be used to distinguish the **3 classes**?

The Inverse Problem

Example 10.3 (*Eggs*)

If we let the random variable X denote the weight of an egg, we need to find the values of x_1 and x_2 indicated in the following diagram,



The Inverse Problem

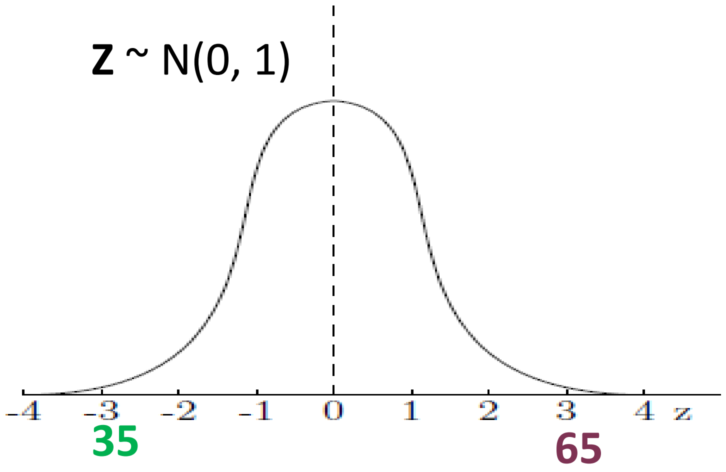
Example 10.3 (*Eggs*)

- Common sense, and our knowledge of the Normal distribution, tells us that the value of x_1 is a little below the mean value of 50g and the value of x_2 somewhat higher than the mean value of 50g.

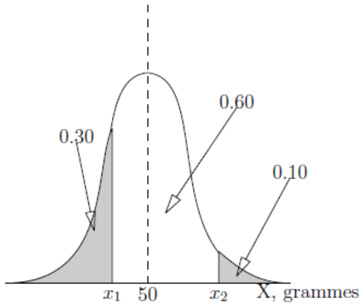
- In fact, we can usually make a reasonably accurate guess if we use the result that **virtually all the curve (99.7%) is contained within the limits ± 3 standard deviations** either side of the mean.

- In this case, virtually all the eggs will lie in the **range**,

$$[50 - 3 \times 5, 50 + 3 \times 5]=[35, 65]\text{g}.$$



The Inverse Problem



Example 10.3 (*Eggs*)

- In order to solve the problem, we have to find the values of a standard normal variable (Z) corresponding to the **same probabilities** indicated on the diagram above.

- The Z **value** exceeded with a **probability 0.1** is found to be **1.2816**, i.e. ***P*(*Z* ≥ 1.2816) = 0.1**. At the other end of the distribution we find the probability a Z **value** is **less than 0.3** is **-0.5244**, i.e. ***P*(*Z* ≤ −0.5244) = 0.3**.

Note that this last value was **found using the symmetry** of the distribution.

- You should check that these are the **z values** you would have obtained if you had been working the other way round.

The Inverse Problem

Example 10.3 (*Eggs*)

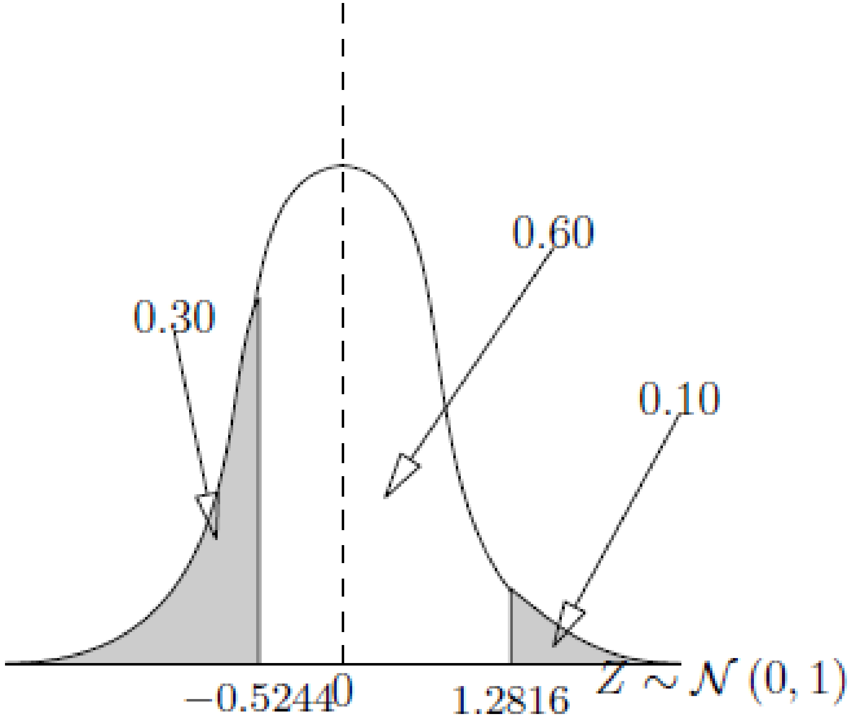
- **Note** that, **if the required probability is not in the inverse tables**, you have to **use the main table backwards**.
- To do this, look for the required probability (**or as close as you can get to it**) in the body of the table, and then read off the corresponding z-value.

*For example, scanning through the body of the table, we find that a probability of **0.1003** corresponds to a **z value** of **1.28** and a probability of **0.0985** corresponds to a **z-value** of **1.29**. Clearly, the **actual value** corresponding to a probability of **0.10** (which we know is **1.2816**) is somewhere **between 1.28 and 1.29**.*

- In practice, a good estimate can be found using **interpolation**.

The Inverse Problem

Example 10.3 (*Eggs*)



The final stage of the problem is to **apply the inverse transformation** to get the appropriate **values on the original scale i.e. *X***.

The Inverse Problem

Example 10.3 (*Eggs*)

Recall, the inverse transformation is, $\boldsymbol{X = \sigma z + \mu}$. Thus we, have,

- The weight which will be exceeded by the **largest 10%** of eggs is given by,

$$X = 5 \times 1.2816 + 50$$
$$= \mathbf{56.41g}$$

- The weight which the **smallest 30%** of eggs will lie below is given by,

$$X = 5 \times -0.5244 + 50$$
$$= \mathbf{47.38g}$$

Finally, a word of warning. It is easy to, and very common, to get the probabilities the wrong way round. Draw a diagram to help you understand the statement of the problem.

Normal Sampling Distributions (Chapter 10)

All the problems considered so far have supposed that a **single measurement** is randomly drawn from some population in which the possible values of the measurement **follow a Normal distribution**.

Now we consider what happens if we,

- Take a **random sample of size n** from a population whose values follow a Normal distribution. We will denote the random sample of values by $X_1, X_2, \dots, X_n$.
  For example, we might randomly select n = 10 people and measure their height.
- Calculate the **mean value** of the sample, i.e. $\boldsymbol{\bar{X} = \sum_i \frac{X_i}{n}}$.
- Calculate the **total value** of the sample, i.e. $\boldsymbol{T = \sum_i X_i}$

Now it should be clear that, since each member of the sample is a random variable, **the sample mean must also be a random variable**.

The question is what is its distribution.

Distribution of the Sample Mean

The following result can only be quoted since its proof is beyond the level of this module.

Theorem 10.1

Suppose that random variables $X_1, X_2, \ldots, X_n$ each follow a Normal distribution with **mean μ** and **variance σ^2**, i.e. $X_i \sim N(\mu, \sigma^2)$. Then, for the **sample mean $\bar{X}$**, we have the ***sampling distribution*** of that ***statistic*** is,

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

Note: *A statistic is any quantity calculated from a sample, e.g. mean, median, minimum, maximum etc.* ***The sampling distribution is just the probability distribution of that statistic.***

Distribution of the Sample Mean

Theorem 10.1

- What this result says is that the **sample mean** ($\bar{X}$) **has the same theoretical population mean**, μ, as any single value drawn from the population, but that **its variance is reduced by a factor** n.

- Given our knowledge of the role of the variance in the Normal distribution, the result suggests that **sample mean** ($\bar{X}$) **ought to lie closer to the true population mean** μ **as the sample size** (n) **increases**.

Distribution of the Sample Mean

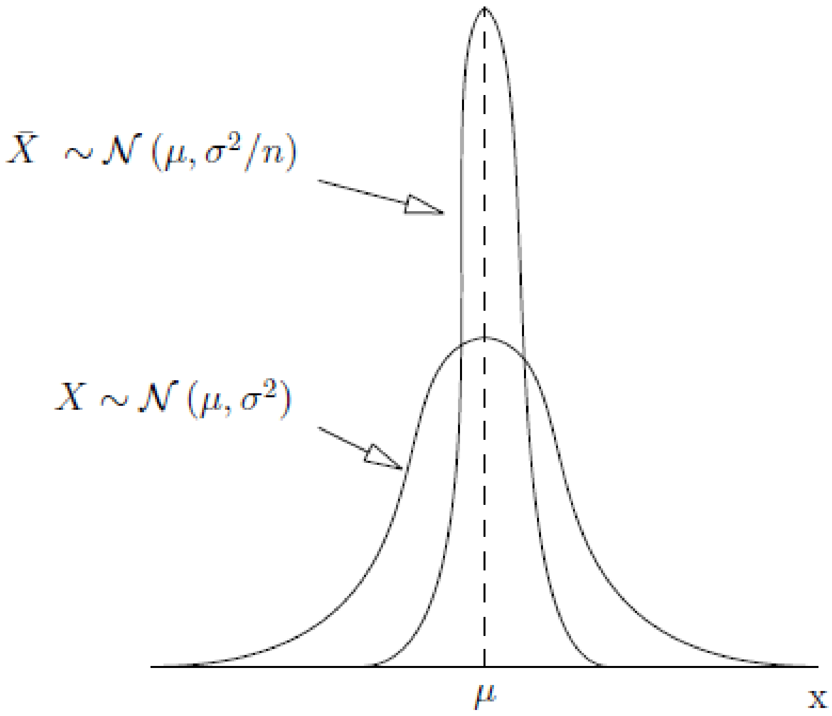
Definition 10.1 (*Standard Error*)

The quantity $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ is called the **standard error of the mean**. It is essentially the **same as the standard deviation** for a single observation but reflects the fact that the variance of the mean depends on the sample size n.

Distribution of the Sample Mean

Definition 10.1 (*Standard Error*)

The result can be represented pictorially as follows,



- **As n gets larger and larger**, the distribution of the **sample mean** gets **more and more concentrated around the value of the population mean μ**.
- In practice, this suggests that, if the true population mean were unknown, the **sample mean ought to be a good estimate** of its value.
- This is a **consequence of the WLLN** (Theorem 6.2) and that the Normal distribution is closed under linear combinations.

Distribution of the Sample Mean

Manchester Metropolitan University

Example 10.1 (*Apple weights*)

In **Example 9.4** we assumed that the weight of individual apples sold by a supermarket were Normally distributed with **mean 100g** and **standard deviation 8g**, i.e. if the random variable $\boldsymbol{X}$ represents the weight then $\boldsymbol{X \sim N(100, 8^2)}$.

The supermarket also sells apples in **cartons of four**. What is the probability that the mean weight of the apples in a **randomly selected carton** is,

1. more than 105g
2. less than 98g
3. between 98 and 102g

Distribution of the Sample Mean

Example 10.1 (*Apple weights*)

Here n = 4, μ = 100 and σ = 8. If we denote the mean weight of the apples in a carton by $\bar{X}$, then,

$$\bar{X} \sim N\left(100, \frac{8^2}{4}\right) \sim N(100,16)$$

$$\boldsymbol{Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)}$$

In this case the **standard error of the mean** is $\frac{8}{\sqrt{4}} = \sqrt{16} = \mathbf{4}$.

Having calculated the **standard error**, we answer such problems in the same way as before by converting the problem to one involving the standard Normal distribution.

Distribution of the Sample Mean

Example 10.1 (*Apple weights*)

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Table Rules:

1. $P(Z > +a)$ - Found directly from table
2. $P(Z < +a) = 1 - P(Z > +a)$ - Complement law
3. $P(a < Z < b) = P(Z \geq a) - P(Z \geq b)$ - Between two points a and b.
4. $P(Z < -a) = P(Z < +a)$ - Symmetry
5. $P(Z > -a) = 1 - P(Z > +a)$

1.

$$P(\bar{X} \geq 105) = P\left(Z \geq \frac{105 - 100}{4}\right)$$

$$= P(Z \geq 1.25)$$

$$= \mathbf{0.1056}$$

2.

$$P(\bar{X} \leq 98) = P\left(Z \geq \frac{98 - 100}{4}\right)$$

$$= P(Z \geq -0.5) = P(Z \geq 0.5) \text{ Symmetry}$$

$$= \mathbf{0.3085}$$

Distribution of the Sample Mean

Example 10.1 (*Apple weights*)

3.

$$P(98 \leq \overline{X} \leq 102) = P\left(\frac{98-100}{4} \leq Z \leq \frac{102-100}{4}\right)$$
$$= P(-0.5 \leq Z \leq 0.5)$$
$$= P(Z \geq -0.5) - P(Z \geq 0.5) \quad \text{(Rule 3)}$$
$$= 1 - P(Z \geq 0.5) - P(Z \geq 0.5) \quad \text{(Rule 5)}$$
$$= 1 - 2 \times P(Z \geq 0.5)$$
$$= 1 - 2 \times 0.3085$$
$$= \mathbf{0.3830}$$

Table Rules:

1. $P(Z > +a)$ - Found directly from table
2. $P(Z < +a) = 1 - P(Z > +a)$ - Complement law
3. $P(a < Z < b) = P(Z \geq a) - P(Z \geq b)$ - Between two points a and b.
4. $P(Z < -a) = P(Z < +a)$ - Symmetry
5. $P(Z > -a) = 1 - P(Z > +a)$

The Central Limit Theorem (CLT)

- This is a remarkable theorem which explains why the Normal distribution plays such an important role in statistics.

The theorem essentially says that, under very mild conditions, the distribution of the sample mean will always tend towards that of a Normal distribution, no matter what the source of the original data.

- Many practical studies involve **calculating the mean of some process**, for example the mean waiting time for a bus.
- The **CLT** says that, even if the actual waiting times follow some other distribution, e.g. an exponential, the mean waiting time will be **approximately Normal in large samples**.

Distribution of the Sample Total

Suppose we have an independent random sample of size n from a Normal distribution, e.g. $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$. Define the random variable $\boldsymbol{T} = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{X_i}$, then,

$$\boldsymbol{T \sim N(n\mu, n\sigma^2)}$$

That is, the *sampling distribution* of the **total**, $\boldsymbol{T}$, is also Normally distributed with $\boldsymbol{E[T]} = \boldsymbol{n\mu}$ and $\mathbf{var}(\boldsymbol{T}) = \boldsymbol{n\sigma^2}$.

$$\boxed{\boldsymbol{Z} = \frac{\boldsymbol{T - \mu}}{\sqrt{\boldsymbol{n}}\boldsymbol{\sigma}} \sim \boldsymbol{N(0,1)}}$$

Distribution of the Sample Mean

Example 10.2 (*Apple weights again*)

In Example 9.4 we assumed that the weight of individual apples sold by a supermarket were Normally distributed with mean 100g and standard deviation 8g, i.e. if the random variable X represents the weight then $X \sim N(100, 8^2)$.

Let the random variable T denote the total weight of a carton of 4 apples.
Find the probability that the total weight of a carton is,

1. more than 450g

2. between 375g and 425g

Distribution of the Sample Mean

Manchester Metropolitan University

Example 10.2 (*Apple weights again*)

Here n = 4, μ = 100 and σ = 8 so that, the total weight,

$$T \sim N(4 \times 100, 4 \times 8^2) \sim N(400, 16^2)$$

We answer the problem by converting to a standard Normal as before,

1. We have,

$$\boldsymbol{Z = \frac{T - \mu}{\sqrt{n}\sigma} \sim N(0, 1)}$$

$$P(T > 450) = P\left(Z > \frac{450 - 400}{16}\right)$$

$$= P(Z > 3.13)$$

$$= \mathbf{0.0009}$$

Distribution of the Sample Mean

Manchester Metropolitan University

$$\boldsymbol{Z = \frac{T - \mu}{\sqrt{n}\sigma} \sim N(0,1)}$$

Example 10.2 (*Apple weights again*)

2. We have,

$$P(375 < T < 425) = P\left(\frac{375 - 400}{16} < Z > \frac{425 - 400}{16}\right)$$

$$= P(-1.56 < Z < 1.56)$$

$$= 1 - 2 \times P(Z > 1.56)$$

$$= 1 - 2 \times 0.0594$$

$$= \mathbf{0.8812}$$