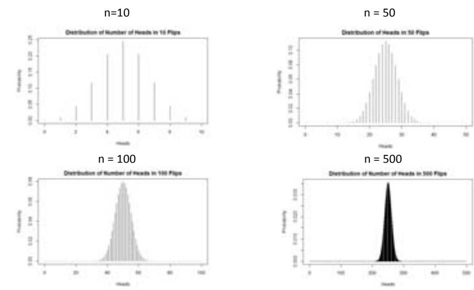
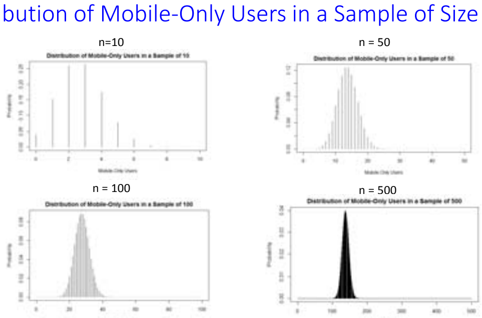
The Normal Distribution

Distribution of Number of Heads in *n* Flips



As n increases, the distribution resembles a bell-shaped distribution centred at the mean of the original distribution.

Distribution of Mobile-Only Users in a Sample of Size *n*



Even though we started with a non-symmetric distribution, as n increases, the distribution resembles a bell-shaped distribution centred at the mean of the original distribution as well. Coincidence?

What features do we see?

- The distributions from these binomial experiments tend to share similar features as the sample size increases:
  - they are symmetric
  - they have a bell-shape
  - they are centred at the same point as the mean of the original distribution
  - The bell-shaped curve becomes narrower as *n* increases
- This is not coincidence, but a consequence of the fact that the binomial distribution (a discrete distribution) can be approximated by the normal distribution (a continuous distribution) as n grows.

The Normal Distribution

- One of the most widely used probability models for continuous numerical variables
- Also called the "bell curve" because of its shape
- It has mean $\boldsymbol{\mu}$.
- It has standard deviation $\boldsymbol{\sigma}$.
- The notation $\boldsymbol{N(\mu, \sigma)}$ represents a Normal distribution with mean μ and standard deviation σ.

Preamble: Probability Distributions for Continuous RVs

- We cannot list all outcomes for a continuous random variable, so we cannot represent the probability distribution using a table.
- The probabilities for a continuous random variable are represented as areas under a density curve.

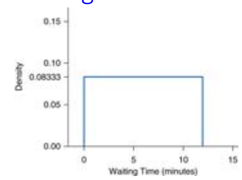
Example: Density Curve

This density curve shows the distribution of waiting times before service at a coffee shop.

The shaded area represents the probability of a waiting time between 0 and 2 minutes.

The total area under the curve is 1.

Example: Finding Probabilities as Areas

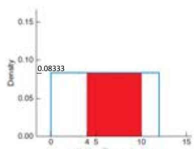


Height is 1/12 so that the area under the curve is 1.

A bus arrives at a certain stop every 12 minutes. The graph shows the density curve for waiting times before the bus arrives. It is a uniform distribution.

Find the probability of a waiting time between 4 and 10 minutes.

Example: Finding Probabilities as Areas



The shaded area represents the probability of a waiting time between 4 and 10 minutes.

We can find the area of this rectangle as length x height:

6 x 1/12 = 1/2

Properties of probabilities for continuous random variables:

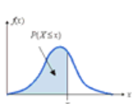
1. The total area under the density curve must equal 1.
2. Every point on the curve must have a vertical height that is 0 or greater.
3. The probability that the random variable takes values between a and b, denoted by $P(a < X < b)$, is the area under the curve between a and b.
4. There is no area under the curve at any particular value, so the probability of a single exact value is 0.
5. Because the probability of single values is 0, all of the following statements are equivalent:

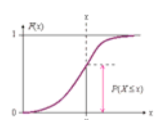
$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b)$$

Finding Areas Under Curves

- We define the **cumulative probability distribution function** of a random variable *X* as

$$F(x) = P\{X \le x\}$$





- Note that for all distributions,

  as x tends to $-\infty$, F(x) tends to 0 (no accumulated probability)

  as x tends to $+\infty$, F(x) tends to 1 (probability of the sample space)

Finding Areas Under Curves (cont.)

- The cumulative distribution function can be used to evaluate probabilities as follows:

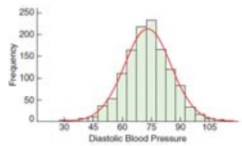
$$P\{a < X \le b\} = F(b) - F(a)$$

$$P\{X \le b\} = F(b)$$

$$P\{X > a\} = 1 - F(a)$$

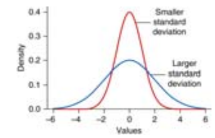
- For all the distributions in this course, we will always find probabilities of intervals using the cumulative distribution function, which is programmed in R.

Visualizing the Normal Model



This histogram shows diastolic blood pressure readings for a sample of males. The distribution can be modelled using a Normal distribution.

The Normal Distribution: Visualizing Centre and Spread

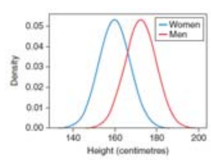


These two Normal distributions have the same mean but different standard deviations.

Note that as with all distributions, the areas under the curves represent the proportions of observations falling in each interval.

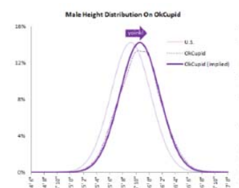
Therefore, the total area under these curves is always equal to 1.

The Normal Distribution: Visualizing Centre and Spread



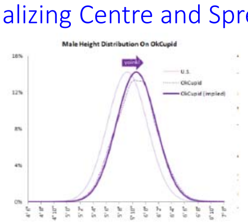
These normal curves show the distribution of heights for adult men and women. Notice that the spread of each curve (the standard deviation) is about the same, but that the centre of each curve (the mean) is different.

The Normal Distribution: Visualizing Centre and Spread



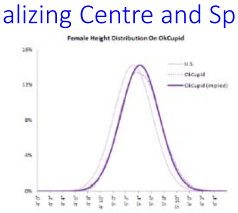
These normal curves show the distribution of heights for adult men in the US and the distribution of heights as reported in the dating site OKCupid. Notice that the curves are almost the same, but that the centre of each curve (the mean) is different: men tend to report their height a couple of inches higher than it actually is!

The Normal Distribution: Visualizing Centre and Spread



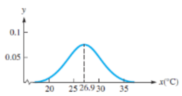
Moreover, starting at roughly 5'8", the top of the dotted curve tilts even further rightward. This means that, as they get closer to six feet, men round up a bit more than usual, stretching for that coveted psychological benchmark.

The Normal Distribution: Visualizing Centre and Spread



Surprisingly, women also tend to report their height a couple of inches higher than it actually is (but without the lurch towards a benchmark height).

The Normal Distribution: Reliability of the Confederation Bridge



- Quantities that were found to be normally distributed (or whose logarithms were found to be normally distributed) arose in the analysis of dead loads, live loads due to vehicles, wind loads, temperature loads, and ice loads.
- One specific example: after analysing records of daily average temperatures in the region for 46 years, it was concluded that the 3-day temperature drop that was equalled or exceeded 100 times in 100 years is distributed normally, with mean 26.9°C and standard deviation 3.2°C.

"Design criteria and load and resistance factors for the Confederation Bridge," by J.G. MacGregor et al. (*Can. J. Civ. Eng.*, 24, 882–897 (1997)

The Normal Distribution: Quality Control

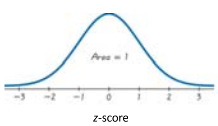
Suppose the volume of sauce in a 700 ml jar of sauce is distributed normally, with standard deviation 5 ml.

Quality control engineers must calibrate their filling machines to a mean of 710.25 ml to ensure that at least 98% of all jars have content of at least 700 ml. (Solution to come.)

Standard Normal Distribution

The standard normal distribution is a normal distribution with $\mu = 0$ and $\sigma = 1$.

Once again, the total area under the curve is equal to 1.



The values for a standard normal distribution are called z-scores

z-Scores

z-score

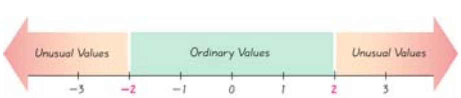
Measures how many standard deviations an observed data value is from the mean.

Example:

A *z*-score of 1.5 means the observed data value is 1.5 standard deviations above the mean.

A *z*-score of –1.5 means the observed data value is 1.5 standard deviations below the mean.

Interpreting *z*-Scores



Whenever a value is less than the mean, its corresponding *z*-score is negative

Ordinary values: $-2 \le z\text{-score} \le 2$

Unusual Values: $z\text{-score} < -2 \text{ or } z\text{-score} > 2$

Using *z*-Scores

We can use z-scores to:

- Compare the relative standing of observations from different groups.
- Identify which observations are more unusual than others.
- Find percentiles for normal distributions that are not necessarily standard.

Example: Making Comparisons with z-scores

Mary scored 70 out of 100 (70%) on her first statistics exam and 33 out of 45 (73%) on her second statistics exam. On the first exam, the mean was 60 and the standard deviation was 10. On the second exam, the mean was 36 and the standard deviation was 3.

On which exam did Mary perform better when compared to the whole class?

Exam	Grade	Mean	Std. Deviation
Exam 1	70/100	60	10
Exam 2	33/45	36	3

We can obtain the relative standing of each observation by finding its z-score.

Computing z-Scores

The formula for calculating the z-score of an observation x in a population with mean μ and standard deviation σ is:

$$z = \frac{x-\mu}{\sigma}$$

Distance from mean

Divide by standard deviation to determine how many standard deviations x is from the mean

The process is called standardizing.

Example: Making Comparisons with z-scores

On which exam did Mary perform better when compared to the whole class?

Exam	Grade	Mean	Std. Dev.
Exam 1	70/100	60	10
Exam 2	33/45	36	3

$\left(z = \frac{x-\mu}{\sigma}\right)$

First exam: $z = \frac{70-60}{10} = 1$

Second exam: $z = \frac{33-36}{3} = -1$

The z-score is higher for the first exam than for the second one, so she did better on the first exam when compared to the rest of the class.

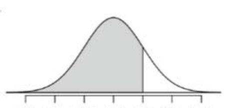
Percentiles

Mary would like to know what percentage of students got a grade lower than hers. That is, she would like to know her percentile.

If the distribution is normal (any mean or standard deviation), we can use the standard normal distribution (mean 0 and standard deviation 1) to work with percentiles.

Probabilities for the Standard Normal

Graphically speaking, the percentile of a given score is the area under the curve to the left of that score

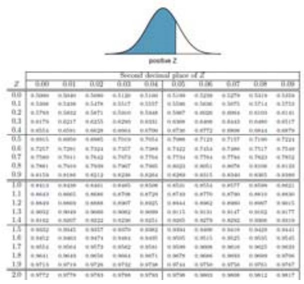


We will learn how to find percentiles using tables, or using technology (R, some calculators, and applets).

Finding Standard Normal Probabilities

- The first and most helpful step is to draw the curve, label it appropriately, and shade in the region of interest.
- Find the percentile of interest using technology or the standard normal table (Table B1 in the Appendix).
- Table B1 uses cumulative areas from the LEFT (the cumulative distribution function).
- Although we will in one example how tables such as Table B1 are used, we will always use technology for calculations regarding the normal distribution.
- Note: percentiles are also interpreted as the probability of observing that value or one lower.

Table B1 (positive values)



Using Table B1

1. It is designed only for the *standard* normal distribution (mean 0, standard deviation 1).
2. Two pages: one for *negative z* scores and one for *positive z* scores.
3. Each value in the body of the table represents a *cumulative area from the left* up to a vertical boundary above a specific *z* score.

Using Table B1

4. Avoid confusion between *z* scores and areas.

   z score: *Distance* along horizontal scale of the standard normal distribution; refer to the <u>leftmost column and top row</u> of Table B1.

   Area: *Region* under the curve; refer to the values in the <u>body</u> of Table B1.

5. The part of the *z* score denoting hundredths is found across the top.

Example – Bone Density Test

A bone mineral density test measures density as a *z*-score, which has a standard normal distribution.

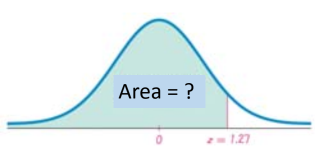
A randomly selected adult undergoes a bone density test.

Find the percentile of a reading of 1.27.

Example – continued

We write this as:

$P(z < 1.27) = ?$



Using Tables (Table B1)

1.27

1.2 on first column

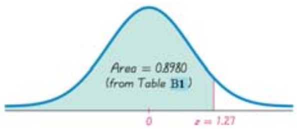
0.07 on first row

z	Second decimal place of 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	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

0.8980 is the number we are after

Example – continued

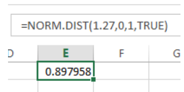
$P(z < 1.27) = 0.8980$



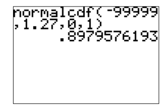
A bone density of 1.27 is the 89.8 percentile of this distribution. This means that 89.8% of observations are below the value of 1.27.

Example: Using Other Technology

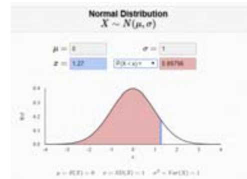
Excel



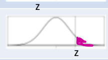
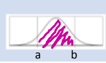
TI Calculators



Online Applets



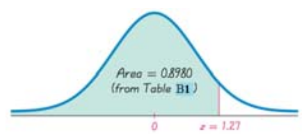
Finding areas for given scores with R: pnorm()

Area of interest	Procedure	Picture
To the left of a single score z	Use the pnorm(z) value obtained directly.	
To the right of a single score z	Use 1 minus the pnorm(z) value obtained.	
Between two scores a and b	Subtract the smaller area from the larger one: pnorm(b)- pnorm(a)	

The instruction **pnorm(z)** gives the area to the left of z under the standard normal distribution.

Example – continued, using R

$P(z < 1.27) = 0.8980$



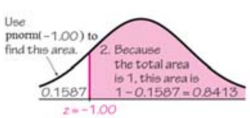
With R:

```
> pnorm(1.27)
[1] 0.8979577
```

A bone density of 1.27 is the 89.8 percentile of this distribution. This means that 89.8% of observations are below the value of 1.27.

Example – Area to the Right

Using the same bone density test, find the percentage of observations that have a result above –1.00 (which is considered to be in the "normal" range of bone density readings).

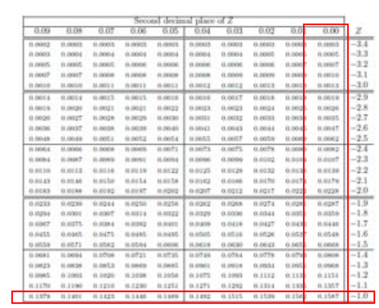



```
> 1-pnorm(-1)
[1] 0.8413447
```

The proportion of observations having a bone density above –1 is 0.8413, so the percentage is 84.13%.

-1.00

With Table B1:

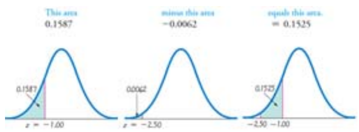


Example – Area Between Two Scores

A bone density reading between –1.00 and –2.50 indicates the subject has osteopenia. Find the percentage of bone density readings that show osteopenia.

```
>
> pnorm(-1)-pnorm(-2.5)
[1] 0.1524456
```

1. The area to the left of $z = -2.50$ is 0.0062.
2. The area to the left of $z = -1.00$ is 0.1587.
3. The area between $z = -2.50$ and $z = -1.00$ is the difference between the areas found above.



Repeating the Summary with Examples

```
> pnorm(-1.39)
[1] 0.08226444
> pnorm(-1.21)
[1] 0.1131394
```

Example	Area of interest	Procedure	Result
$P(z < -1.39)$	To the left of a single *z*-score	Use the value pnorm(z) directly.	$P(z < -1.39) = 0.0823$
$P(z > -1.21)$	To the right of a single *z*-score	Obtain 1 minus the value pnorm(z)	$P(z > -1.21) = 1 - P(z < -1.21)$ $= 1 - 0.1131 = 0.8869$
$P(-1.39 < z < -1.21)$	Between two z-scores	Subtract the smaller area from the larger one	$P(-1.39 < z < -1.21)$ $= 0.1131 - 0.0823 = 0.0308$

If *z* is a standard normal variable, find the probability that *z* lies between 0.7 and 1.98.

A. 0.7580

B. –0.2181

C. 0.9761

D. 0.2181

```
> pnorm(0.7)
[1] 0.7580363
> pnorm(1.98)
[1] 0.9761482
```

Finding *z*-Scores for Given Percentiles Using Tables

1. Draw a bell-shaped curve and identify the region under the curve that corresponds to the given percentile of interest.
2. If using Table B1, locate the closest proportion in the body of the table and identify the corresponding *z* score.

Finding *z*-Scores for Given Percentiles Using R

1. Draw a bell-shaped curve and identify the region under the curve that corresponds to the given percentile of interest.
2. Use the **inverse normal** function applied to the percentile p. This function is

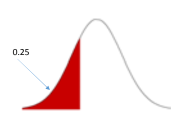
qnorm(p)

In other words, **qnorm(p)** returns the score value z that leaves an area of p to its left under the standard normal distribution

Example: Score for a Given Percentile

Find the value of the 25th percentile of the standard normal.

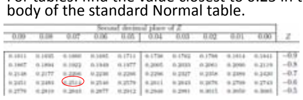
In other words, find the z-score such that 25% of the values on the distribution are below it.



Example (continued)

```
> qnorm(0.25)
[1] -0.6744898
```

For tables: find the value closest to 0.25 in the body of the standard Normal table.



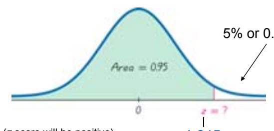
Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	z
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	−0.9
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	−0.8
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	−0.7
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	−0.6
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	−0.5

Here, it is 0.2514.

The z-score corresponding to this value is -0.67.

The 25th percentile is $z = -0.67$.

Finding z-Scores for Given Percentiles



Finding the 95th Percentile

```
> qnorm(0.95)
[1] 1.644854
```

Finding probabilities and scores for general normal distributions

Use z-scores if using tables:

$$z = \frac{x - \mu}{\sigma}$$

After standardizing, the normal distribution has mean 0 and standard deviation 1 (standard).

In R it is not necessary to standardize

- The **pnorm()** command can incorporate the mean and standard deviation of the normal distribution we want, and allows us to skip the standardization process.
- For example, to find the area to the left of 54 in a normal distribution with mean 60 and standard deviation 6, the instruction is:

pnorm(54, mean = 60, sd = 6)

54 tells it the score of interest

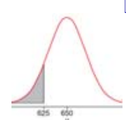
mean is the mean (60 here, and the default is 0)

sd is the standard deviation (6 here, and the default is 1)

Example

Suppose the weight of a grizzly bear cub follows a Normal distribution with a mean weight of 650 grams and a standard deviation of 20 grams. What percentage of newborn grizzly bear cubs weighs less than 625 grams?

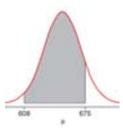
```
> pnorm(625,mean=650,sd=20)
[1] 0.1056498
```



The proportion of newborn grizzly bear cubs that weigh less than 625 grams is 0.1056, or 10.56% of them.

Example

Suppose the weight of a grizzly bear cub follows a Normal distribution with a mean weight of 650 grams and a standard deviation of 20 grams. What is the probability that a newborn grizzly bear cub weighs between 608 and 675 grams?



Find the probabilities for 675 and 608 using the pnorm() instruction with the appropriate mean and sd, and subtract the smaller area from the larger area to obtain the probability:

```
> pnorm(675, mean=650, sd=20)-pnorm(608,mean=650,sd=20)
[1] 0.8764858
```

The probability that a newborn grizzly bear cub weighs between 608 and 675 grams is 0.8765.

Finding Measurements from Areas

(We will call this an inverse Normal problem.)

R command:

qnorm(p,mean,sd)

We only worked with standard normal distributions before (default mean=0, sd=1). The instruction is the same, adding the values of mean and sd.

Normal or Inverse Normal?

It is known that the heights of adult men follow a Normal model. For each of the following situations, decide if it is a Normal or an inverse Normal problem.

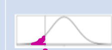
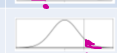
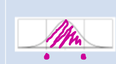
- A clothing store manager wonders what percentage of her customers are taller than 1.90 metres.

  This is a Normal problem.

- A clothing store manager wants to cater to the tallest 20% of men and wants to know what heights she should accommodate.

  This is an inverse Normal problem.

Finding scores for given areas

Score of interest	Procedure	Picture
Area given is to the left of the score of interest	Use qnorm of the area given.	
Area given is to the right of the score of interest	Use qnorm of 1 minus the area given.	
Between two symmetric scores $(-z_0, z_0)$	Subtract the area given from 1 to get the area of the tails; divide by 2 to get the area of one tail. Use qnorm of that area to get the negative of z_0	

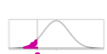
Practice

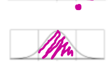
The *N*(160, 7.5) model is a good description of the distribution of women's heights (in centimetres).

Find the 25th percentile of this distribution.

In other words, find a value, in centimetres, such that 25% of the values on the distribution are below it.

Score of interest	Procedure
Area given is to the left of the score of interest	Use qnorm of the area given.



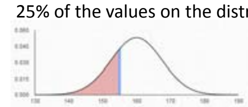


Practice

The *N*(160, 7.5) model is a good description of the distribution of women's heights (in centimetres).

Find the 25th percentile of this distribution.

In other words, find a value, in centimetres, such that 25% of the values on the distribution are below it.



```
> qnorm(0.25,160,7.5)
[1] 154.9413
> qnorm(0.75,160,7.5)
[1] 165.0587
> pnorm(25,160.7,7.5)
[1] 1.798301e-73
```

A) 154.9 B) 165.1 C) 25 D) 180