Chapter 6
Continuous Random Variables

Aims

In this chapter we deal with:

Continuous random variables

Probability density functions

Normal distributions

Working in standard units (i.e., *z*-scores)

By the end of this chapter you should:

- know the properties of a probability density function
- know the properties of a Normal probability density function
    - in particular, the 68–95–99.7 rule
- be able to use software and computer output to solve
    - Normal probability problems
    - Inverse Normal problems
- understand and be able to apply *z*-scores

Problem: Diagnosing Spina Bifida

A screening test for spina bifida in a foetus involves measuring the concentration of alpha fetoprotein in the mother's urine.

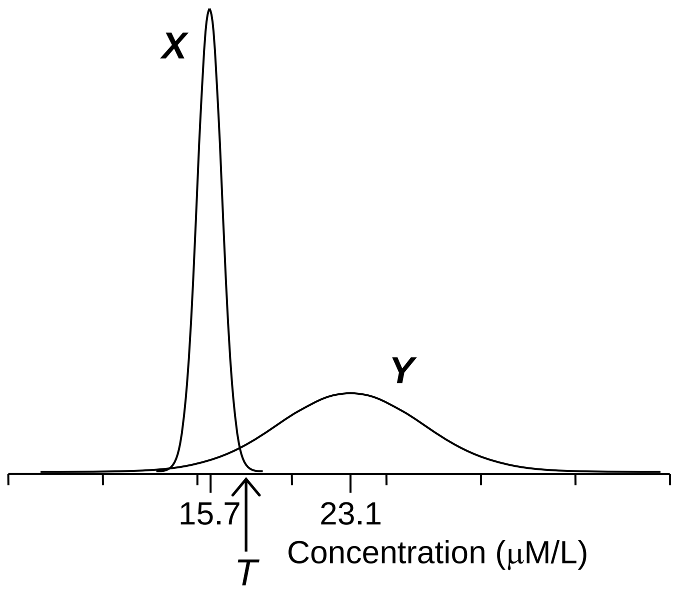
For mothers carrying healthy foetuses, the mean is 15.7 micromoles/litre (μM/L) and the standard deviation is 0.7 μM/L. For mothers carrying foetuses with spina bifida, the mean is 23.1 μM/L and the standard deviation is 4.1 μM/L.

To operate this screening test for spina bifida, medical professionals must set a threshold concentration of alpha fetoprotein, T, say. If the alpha fetoprotein concentration is below T, the foetus is diagnosed as not having spina bifida, whereas if it is above T further testing is required.

Let:

X be the concentration of alpha fetoprotein for a mother carrying a healthy foetus

Y be the concentration of alpha fetoprotein for a mother carrying a foetus with spina bifida.



Suppose T was set at 17.8 μM/L:

- What is the probability that a foetus with spina bifida is correctly diagnosed?
- What is the probability that a foetus without spina bifida is correctly diagnosed?

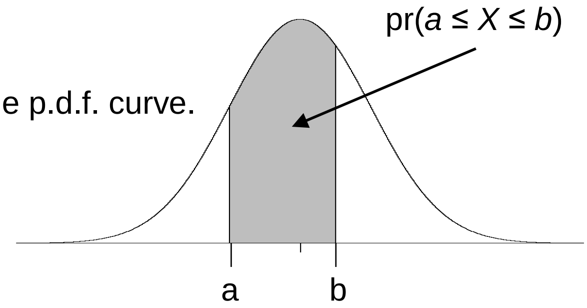
To ensure that 99% of foetuses with spina bifida are correctly diagnosed, at what value should T be set?

Properties of a Probability Density Function (p.d.f.)

1. The p.d.f. curve is always above or on the x-axis.

2. Probabilities are represented by **areas** under the p.d.f. curve.

   pr($a \le X \le b$) = **area** under the p.d.f. curve between $x = a$ and $x = b$.



3. The total area under a p.d.f. curve = **1**

Endpoints of Intervals

For continuous random variables: $\text{pr}(a \le X \le b) = \text{pr}(a < X \le b)$

$= \text{pr}(a \le X < b)$

$= \text{pr}(a < X < b)$

In calculations involving continuous random variables, we do not have to worry about whether interval endpoints are **included (≥ or ≤) or excluded (> or <).**

The Normal Distribution

Heights of Male Students

In a recent survey of 544 male Introductory Statistics students the mean height was 177.1cm and the standard deviation (sd) was 7.7cm.

In this survey:

mean – 1 sd = 177.1cm – 7.7cm = 169.4cm

mean + 1 sd = 177.1cm + 7.7cm = 184.8cm

375/544 = 69% have a height within **1** standard deviation of the mean.

mean – 2 sd = 177.1cm – 2 × 7.7cm = 161.7cm

mean + 2 sd = 177.1cm + 2 × 7.7cm = 192.5cm

515/544 = 95% have a height within **2** standard deviations of the mean.

mean – 3 sd = 177.1cm – 3 × 7.7cm = 154.0cm

mean + 3 sd = 177.1cm + 3 × 7.7cm = 200.2cm

542 / 544 = 99.6% have a height within **3** standard deviations of the mean.

68–95–99.7 Rule

In a Normal distribution, approximately:

68% of observations are within **1** standard deviation of the mean
95% of observations are within **2** standard deviations of the mean
99.7% of observations are within **3** standard deviations of the mean

Percentage Body Fat for Competitive Cyclists

The distribution of percentage body fat for competitive cyclists is modelled by a Normal distribution with a mean of 9 and a standard deviation of 3.

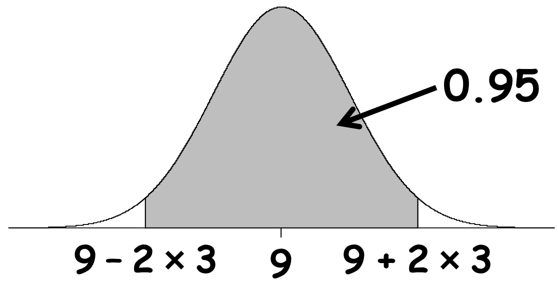
Let X be the percentage body fat of a competitive cyclist.

So $X \sim \text{Normal}(\mu = 9, \sigma = 3)$

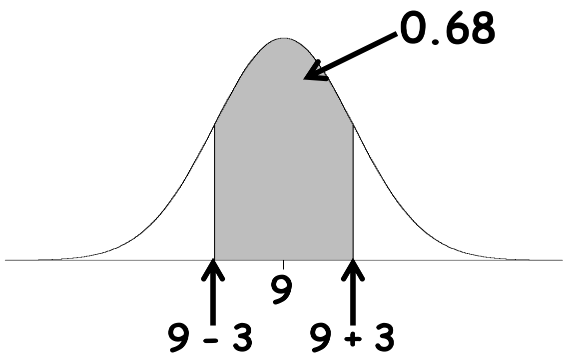
Based on this model:

a. approximately 95% of competitive cyclists have percentage body fat somewhere between

3% and 15%



b. for a randomly chosen competitive cyclist, there is a probability of 0.68 that the cyclist has percentage body fat somewhere between **6% and 12%**



Annual Compound Share Returns for Large Companies

The distribution of annual compound share returns for large companies in the United States (as a percentage) is modelled by a Normal distribution with a mean of 11.0 and a standard deviation of 20.3.

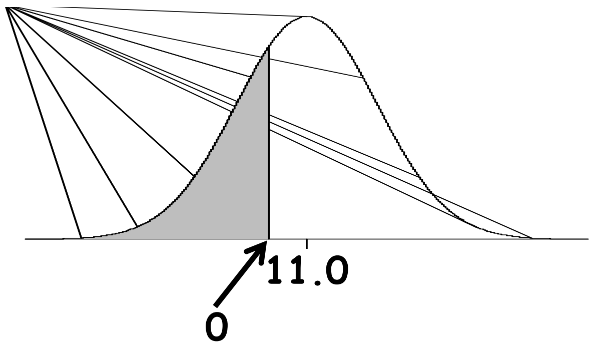
Let X be the annual compound share return for a large company in the United States (as a percentage).

So $X \sim \text{Normal}(\mu = 11.0, \sigma = 20.3)$

Using the computer output given, find the probability that a randomly chosen large US company has an annual compound share return of:

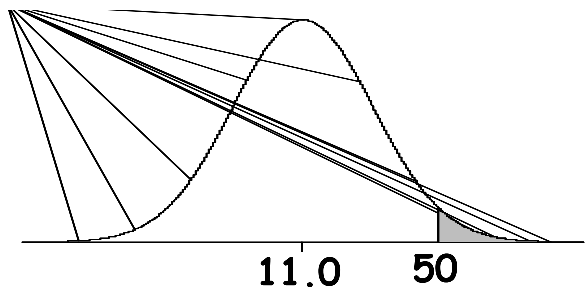
a. less than 0% (i.e., a negative return).

$$\text{pr}(X < 0) = 0.2940$$



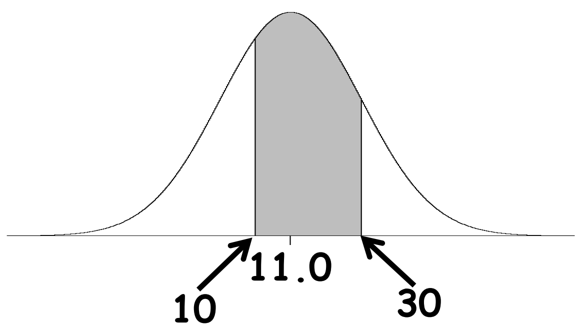
b. more than 50%.

$$\begin{aligned}\text{pr}(X > 50) &= 1 - \text{pr}(X < 50)\\ &= 1 - 0.9726\\ &= 0.0274\end{aligned}$$



c. between 10% and 30%.

$$\begin{aligned}&\text{pr}(10 < X < 30)\\ &= \text{pr}(X < 30) - \text{pr}(X < 10)\\ &= 0.8254 - 0.4804\\ &= 0.3450\end{aligned}$$



Computer Output

```
Normal with mean = 11.0 and standard deviation = 20.3
```

x	P(X <= x)
0	0.2940
10	0.4804
30	0.8254
50	0.9726

Spina Bifida

We will model the distributions of alpha fetoprotein with Normal distributions.

Let:

X be the concentration of alpha fetoprotein for a mother carrying a healthy foetus

Y be the concentration of alpha fetoprotein for a mother carrying a foetus with spina bifida

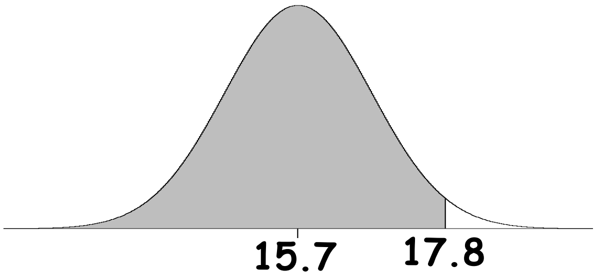
So $X \sim \text{Normal}(\mu_X = 15.7, \sigma_X = 0.7)$

$Y \sim \text{Normal}(\mu_Y = 23.1, \sigma_Y = 4.1)$

Note: The threshold is set at 17.8 μM/L, i.e., if the concentration of alpha fetoprotein is below 17.8 μM/L the foetus would be diagnosed as not having spina bifida.

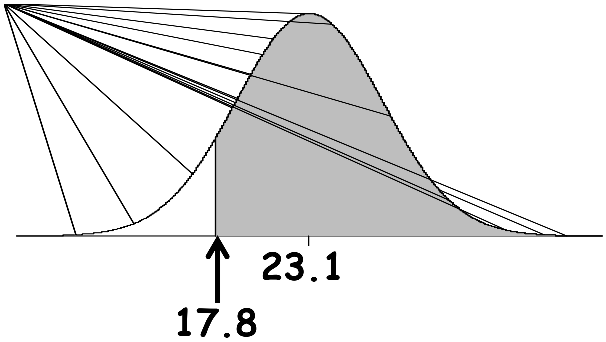
a. For a foetus without spina bifida, what is the probability that it is correctly diagnosed?

pr(*X* < 17.8) = 0.9987



b. What is the probability that a foetus with spina bifida is correctly diagnosed?

$$\text{pr}(Y > 17.8) = 1 - \text{pr}(Y < 17.8) = 1 - 0.0981 = 0.9019$$



Computer Output

```
Normal with mean = 15.7 and
standard deviation = 0.7

     x      P(X <= x)
  17.8       0.9987
```

```
Normal with mean = 23.1 and
standard deviation = 4.1

     x      P(X <= x)
  17.8       0.0981
```

Inverse Normal Problems

IQ Scores

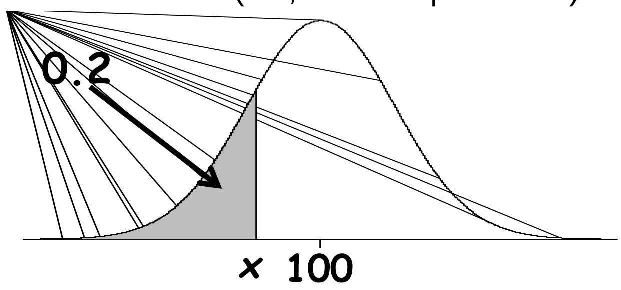
The distribution of IQ scores is modelled by a Normal distribution with a mean of 100 and a standard deviation of 15.

Let X be the IQ of a person.

So X ~ Normal(μ = 100, σ = 15)

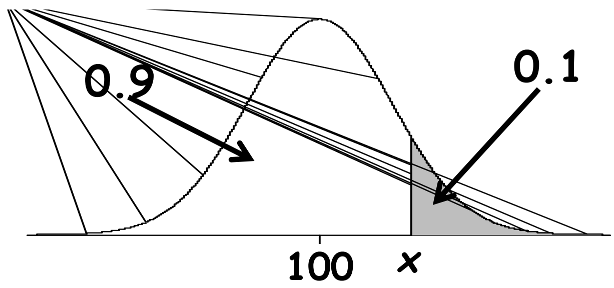
a. Find the IQ score that the bottom 20% of the population fall below (i.e., the 20th percentile).

$$\begin{aligned} pr(X < x) &= 0.2 \\ x &= 87 \end{aligned}$$



b. What IQ score is exceeded by only the top 10% of the population?

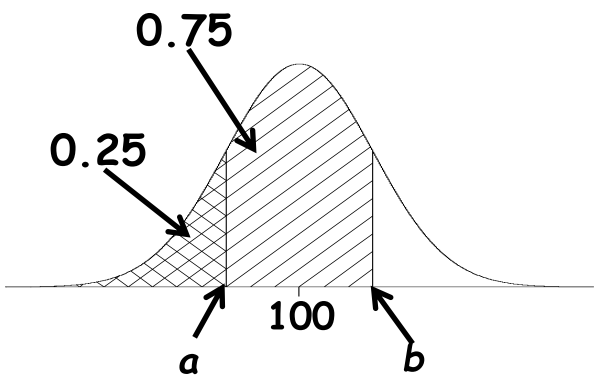
$$\begin{aligned} pr(X > x) &= 0.1 \\ pr(X < x) &= 0.9 \\ x &= 119 \end{aligned}$$



c. Find the interquartile range for IQ scores.

$$\begin{aligned} pr(X < a) &= 0.25 \qquad a = 89.88 \\ pr(X < b) &= 0.75 \qquad b = 110.12 \end{aligned}$$

$$IQR = 110.12 - 89.88 = 20$$



`Computer Output`

```
Normal with mean = 100 and standard deviation = 15

              P(X <= x)        x
                 0.10        80.78
                 0.20        87.38
                 0.25        89.88
                 0.75       110.12
                 0.80       112.62
                 0.90       119.22
```

Human Gestation Period

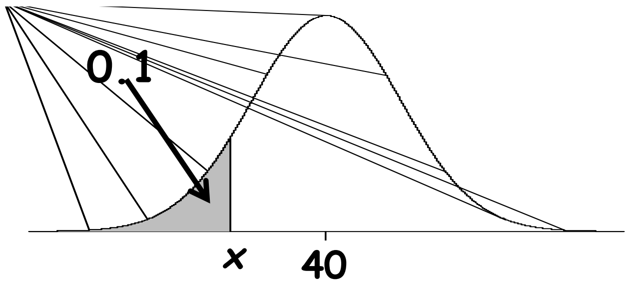
The distribution of the natural human gestation period (in weeks) is modelled by a Normal distribution with a mean of 40 and a standard deviation of 2.3.

Let X be the natural human gestation period in weeks.

So $X \sim \text{Normal}(\mu = 40, \sigma = 2.3)$

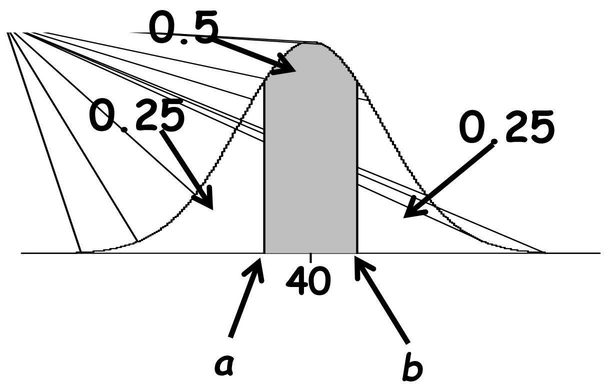
a. Before how many weeks do 10% of births occur?

$$\text{pr}(X < x) = 0.1$$
$$x = 37 \text{ weeks}$$



b. What is the range of gestation period for the central 50% of births?

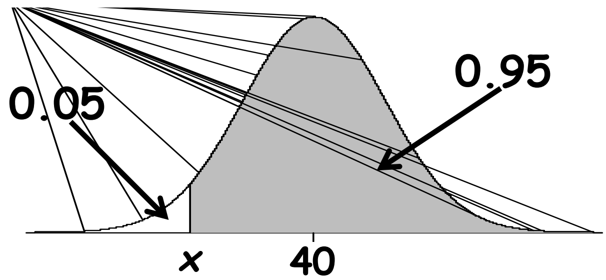
$$\text{pr}(X < a) = 0.25$$
$$a = 38.5$$
$$\text{pr}(X < b) = 0.75$$
$$b = 41.5$$



Range is **38.5** to **41.5** weeks

c. 95% of births occur after how many weeks?

$$\text{pr}(X > x) = 0.95$$
$$\text{pr}(X < x) = 0.05$$
$$x = 36 \text{ weeks}$$



Computer Output

```
Normal with mean = 40 and standard deviation = 2.3

              P(X <= x)        x
                 0.05       36.22
                 0.10       37.05
                 0.25       38.45
                 0.75       41.55
                 0.90       42.95
                 0.95       43.78
```

Working in Standard Units

Test and Examination Results

In a recent semester a student obtained the following marks for the Introductory Statistics mid-semester test and exam.

Test: 19/25

Exam: 40/50

In which of these assessments did this student achieve better results in terms of ranking with students in the same course?

The distribution of test marks is modelled by a Normal distribution with mean 13.5 and standard deviation 4.6 and the distribution of exam marks is modelled by a Normal distribution with mean 30.1 and standard deviation 9.4.

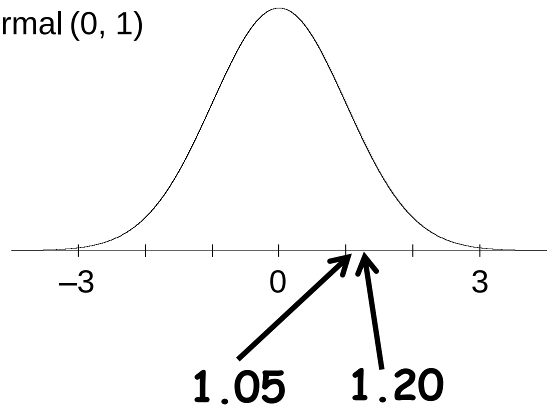
Let T be the test mark
E be the exam mark

So $T \sim \text{Normal}(\mu_T = 13.5, \sigma_T = 4.6)$
$E \sim \text{Normal}(\mu_E = 30.1, \sigma_E = 9.4)$

Test: z-score for 19 = $\dfrac{19 - 13.5}{4.6} = 1.20$

Exam: z-score for 40 = $\dfrac{40 - 30.1}{9.4} = 1.05$

$Z \sim \text{Normal}\,(0, 1)$



This student did better in the **test** in terms of ranking with students in the same course.

Percentage Body Fat for Competitive Swimmers

The distribution of percentage body fat for competitive swimmers is modelled by a Normal distribution with a mean of 10 and a standard deviation of 4.

Let X be the percentage body fat of a competitive swimmer.

So $X \sim \text{Normal}(\mu = 10, \sigma = 4)$

Find the z-score for these percentage body fat values for competitive swimmers:

a. $x = 18$ $z = 2$ (i.e., 18 is 2 sd above the mean of 10)

b. $x = 6$ $z = -1$ (i.e., 6 is 1 sd below the mean of 10)

c. $x = 12.6$ $z =$? (i.e., 12.6 is ? sd above the mean of 10)

The z-score for an observation x:

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

Find the z-score for these percentage body fat values for competitive swimmers:

d. $x = 15.7$ $z = \dfrac{15.7 - 10}{4} = 1.425$

e. $x = 8.9$ $z = \dfrac{8.9 - 10}{4} = -0.275$

Calculate the percentage body fat values (i.e., x-scores) for competitive swimmers with the following z-scores:

f. $z = 2.3$

x is 2.3 sd above the mean so

$x = 10 + 2.3 \times 4 = 19.2$

g. $z = -1.8$

x is 1.8 sd below the mean so

$x = 10 - 1.8 \times 4 = 2.8$

Continuous Random Variables

If a random variable, X, can take any value in some interval it is called a **continuous random variable**.

There are no gaps between the values a continuous random variable can take.

Continuous random variables measure characteristics of items chosen from a:

- clearly defined finite population, or
- random process producing observations

Examples: time, weight, concentration of alpha fetoprotein, annual share return for a company.

Properties of a Probability Density Function (p.d.f.)

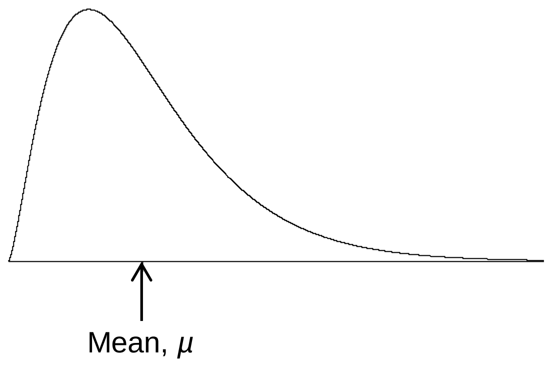
1. The p.d.f. curve is always above or on the x-axis.
2. Probabilities are represented by areas under the p.d.f. curve.

   pr($a \le X \le b$) = area under the p.d.f. curve between $x = a$ and $x = b$.
3. The total area under a p.d.f. curve = 1

Mean (Expected Value) and Standard Deviation

The **expected value** of a random variable X, E(X), is:

- the long-run average for X-values
- also called the **population mean**, μ_X (often shortened to μ)
- where the probability density curve balances



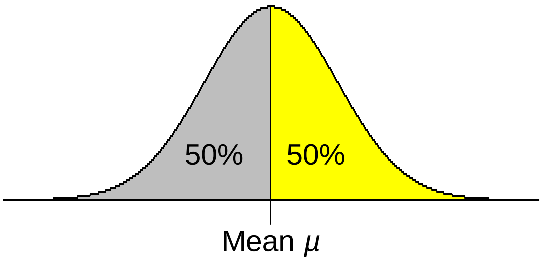
The **standard deviation** of a random variable X, sd(X), is:

- a measure of the variability (spread) of X-values
- roughly, the average distance of the X-values from the population mean
- also called the **population standard deviation**, σ_X (often shortened to σ)

The Normal Distribution

If the distribution of a random variable, X, has a Normal distribution with mean μ and standard deviation σ we write:

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



μ and σ are called the **parameters** of the distribution.

Features of the Normal density curve:

- Symmetric and bell-shaped
- Centred at μ
- σ determines the spread (and hence the height)

68–95–99.7 Rule

In a Normal distribution with mean μ and standard deviation σ, approximately:

68% of observations are within **1** standard deviation of the mean,
i.e., between $\mu - 1\sigma$ and $\mu + 1\sigma$ (or $\mu \pm 1\sigma$)

95% of observations are within **2** standard deviations of the mean,
i.e., between $\mu \pm 2\sigma$

99.7% of observations are within **3** standard deviations of the mean,
i.e., between $\mu \pm 3\sigma$

Obtaining Normal Probabilities

Use statistical software, e.g. Excel, SPSS.

When obtaining Normal probabilities we give the software an ***x*-value** and it returns a **probability**.

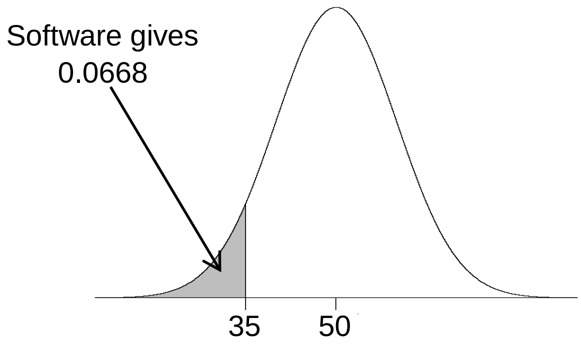
Most software calculates the value of pr($X \le x$), the cumulative or lower-tail probability.

Method for obtaining Normal probabilities:

1. Sketch a Normal curve, marking on the mean and value(s) of interest.
2. Shade the area under the curve corresponding to the required probabilities.
3. Obtain the desired probabilities from the lower-tail probabilities provided by the software.

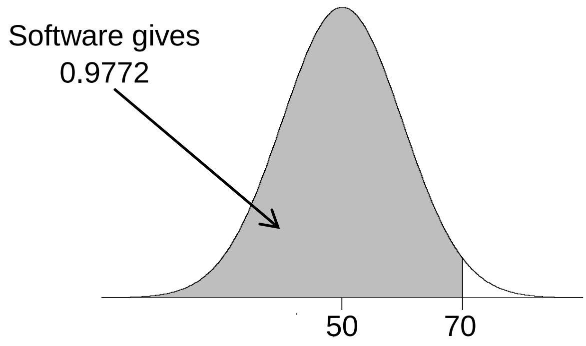
X ~ Normal(μ = 50, σ = 10)

Find pr($X \le 35$)



pr($X \le 35$) = 0.0668

Find pr($X \le 70$)



pr($X \le 70$) = 0.9772

Inverse Normal Problems

Also use statistical software, e.g. Excel, SPSS.

When solving inverse Normal problems we give the software a **probability** and it returns an ***x*-value**.

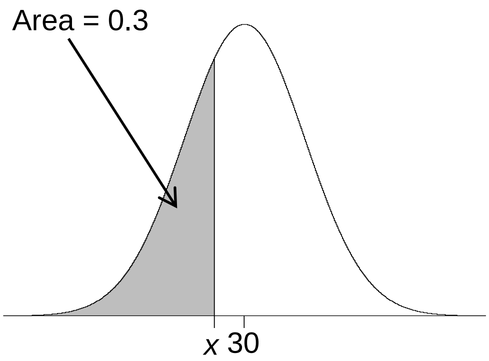
Most software requires the value of pr($X \le x$), the cumulative or lower-tail probability, to be given.

Method for solving inverse Normal problems:

1. Sketch a Normal curve, marking on the mean.
2. Shade the area under the curve corresponding to the given probability.
3. Obtain the desired *x*-value from the lower-tail probability given to the software.

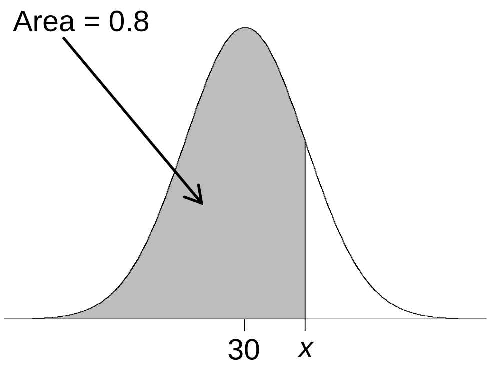
$X \sim$ Normal($\mu = 30$, $\sigma = 5$)

Find x, where pr($X \le x$) = 0.3



Software gives
$x = 27.38$

Find x, where pr($X \le x$) = 0.8



Software gives
$x = 34.21$

Standard Units (*z*-scores)

The ***z*-score** for an observation, x, is the number of standard deviations x is from the mean, μ.

If x is an observation from a Normal distribution with mean μ and standard deviation σ then the ***z*-score** is

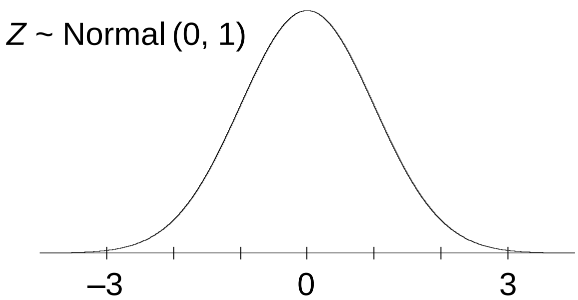
$$\boldsymbol{z = \frac{x - \mu}{\sigma}}$$

- If x is above the mean then the z-score is positive.
- If x is below the mean then the z-score is negative.

Standard Normal Distribution

If $X \sim$ Normal(μ, σ) and $Z = \dfrac{X - \mu}{\sigma}$ then $Z \sim$ Normal(0, 1).

$Z \sim$ Normal(0, 1) is called the **standard Normal distribution**.



Sample Exam / Cecil Test Questions

Questions 1 to **8** refer to the following information.

For mothers carrying foetuses with spina bifida, the distribution of the concentration of alpha fetoprotein (μM/L) in a mother's urine is modelled by a Normal distribution with mean 23.1 and standard deviation 4.1.

Use the following computer output to answer **Questions 1** to **6**.

```
Normal with mean = 23.1 and standard deviation = 4.1

 x   P(X <= x)          P(X <= x)      x
17     0.0684             0.0500    16.356
18     0.1068             0.1000    17.846
19     0.1587             0.2000    19.649
20     0.2248             0.2248    20.000
21     0.3043             0.2400    20.204
22     0.3942             0.2500    20.335
23     0.4903             0.2600    20.462
31     0.9730             0.6785    25.000
32     0.9850             0.7500    25.865
33     0.9921             0.8000    26.551
                          0.9000    28.354
                          0.9500    29.844
```

1. For a randomly selected mother carrying a foetus with spina bifida, the probability that the concentration of alpha fetoprotein in her urine is less than 20.0 μM/L is approximately:

a. 0.1587

b. 0.6957

c. 0.7752

d. 0.2248

e. 0.3043

2. For mothers carrying foetuses with spina bifida, the proportion who have an alpha fetoprotein concentration greater than 18.0 μM/L in their urine is approximately:

a. 0.8413

b. 0.1068

c. 0.8932

d. 0.1587

e. 0.9316

3. For mothers carrying foetuses with spina bifida, the proportion who have an alpha fetoprotein concentration between 22.0 μM/L and 32.0 μM/L in their urine is approximately:

   a. 0.4827

   b. 0.4947

   c. 0.6878

   d. 0.5788

   e. 0.5908

4. For mothers carrying foetuses with spina bifida, the concentration of alpha fetoprotein in urine for which 25% lie below is approximately:

   a. 20.335 μM/L

   b. 25.865 μM/L

   c. 20.204 μM/L

   d. 0.679 μM/L

   e. 20.462 μM/L

5. For mothers carrying foetuses with spina bifida, the concentration of alpha fetoprotein in urine for which 20% lie above is approximately:

   a. 19.649 μM/L

   b. 0.775 μM/L

   c. 28.354 μM/L

   d. 26.551 μM/L

   e. 0.225 μM/L

6. For mothers carrying foetuses with spina bifida, the range of the central 90% of concentrations of alpha fetoprotein in urine is approximately between:

   a. 19.6 μM/L and 26.6 μM/L

   b. 16.4 μM/L and 29.8 μM/L

   c. 17.8 μM/L and 29.8 μM/L

   d. 16.4 μM/L and 28.4 μM/L

   e. 17.8 μM/L and 28.4 μM/L

7. Which **one** of the following statements about mothers carrying foetuses with spina bifida is **false**?

a. Approximately 68% have alpha fetoprotein concentrations between 19.0 μM/L and 27.2 μM/L.

b. Approximately 95% have alpha fetoprotein concentrations between 14.9 μM/L and 31.3 μM/L.

c. Approximately 16% have alpha fetoprotein concentrations greater than 27.2 μM/L.

d. More than 5% have alpha fetoprotein concentrations less than 14.9 μM/L.

e. Almost all have alpha fetoprotein concentrations between 10.8 μM/L and 35.4 μM/L.

8. A mother who is carrying a foetus with spina bifida has an alpha fetoprotein concentration of 16.8 μM/L. What is the *z*-score for this observed concentration?

a. 1.54

b. 11.17

c. –0.55

d. –1.54

e. 0.55

Questions 9 and **10** refer to the following information.

Hypholoma Capnoides is a pleasant tasting mushroom which looks very much like the generally taller, poisonous fungus *Sulphur Tuft*. Let H be the height (in centimetres) of a *Hypholoma Capnoides* mushroom and let S be the height (in centimetres) of a *Sulphur Tuft* fungus. The distribution of H is modelled by a Normal distribution with mean 6.5 and standard deviation 1.76 and the distribution of S is modelled by a Normal distribution with mean 9.5 and standard deviation 1.25.

In summary, $H \sim \text{Normal}(\mu_H = 6.5, \sigma_H = 1.76)$ and $S \sim \text{Normal}(\mu_S = 9.5, \sigma_S = 1.25)$.

9. Which **one** of the following statements is **false**?

The proportion of *Hypholoma Capnoides* mushrooms that are:

a. taller than 11.0 cm is less than the proportion of *Sulphur Tuft* fungi that are taller than 12.5 cm.

b. taller than 5.0 cm is less than the proportion of *Sulphur Tuft* fungi that are taller than 8.0 cm.

c. taller than 9.5 cm is greater than the proportion of *Sulphur Tuft* fungi that are shorter than 6.5 cm.

d. shorter than 8.0 cm is less than the proportion of *Sulphur Tuft* fungi that are taller than 8.5 cm.

e. taller than 4.0 cm is greater than the proportion of *Sulphur Tuft* fungi that are shorter than 11.0 cm.

10. Which **one** of the following statements is **false**?

a. About 2.5% of *Sulphur Tuft* fungi are taller than 12.0 cm.

b. The proportion of *Hypholoma Capnoides* mushrooms that are taller than 10.5 cm is greater than the proportion of *Sulphur Tuft* fungi that are taller than 12.5 cm.

c. About 16% of *Hypholoma Capnoides* mushrooms are shorter than 4.74 cm.

d. *Hypholoma Capnoides* mushrooms are generally shorter than and more variable in height than *Sulphur Tuft* fungi.

e. The proportion of *Hypholoma Capnoides* mushrooms that are shorter than 2.5 cm is less than the proportion of *Sulphur Tuft* fungi that are shorter than 6.0 cm.
