

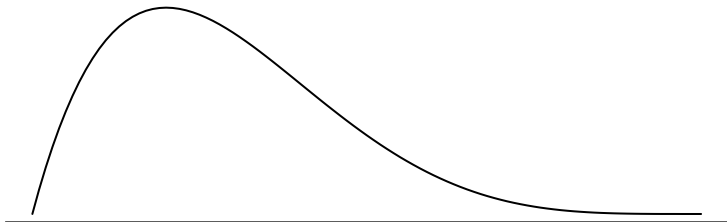
Normal Distributions

Prof Hitchman

9/20/2021

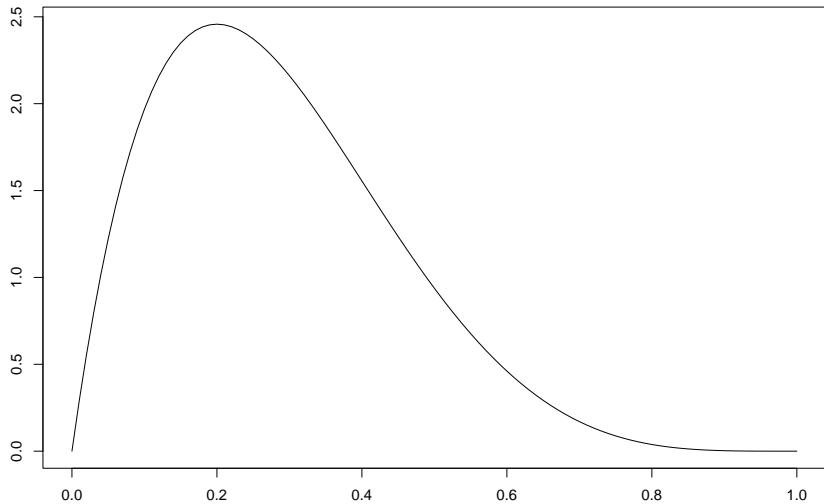
Density Curves

Density Curves

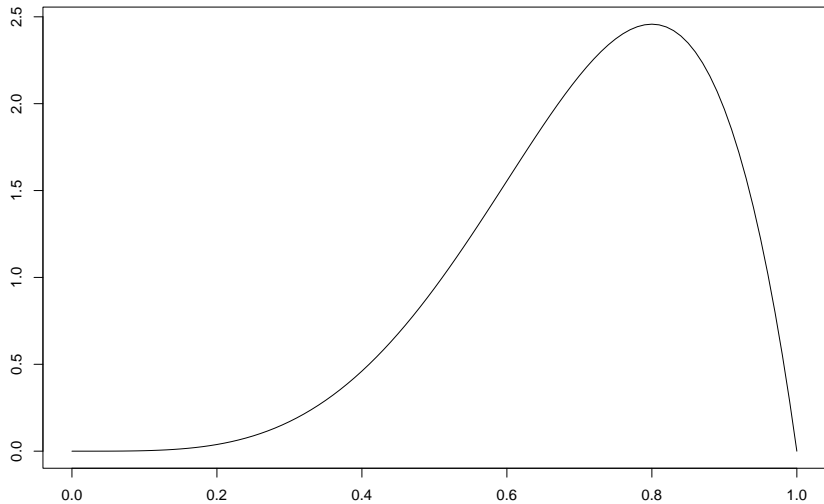


- ▶ Idealized shape of a distribution
- ▶ Two defining features of a density curve:
 - ▶ The area underneath the curve equals 1
 - ▶ The curve is on or above the horizontal axis.
- ▶ With these features in place, area under curve \leftrightarrow proportions

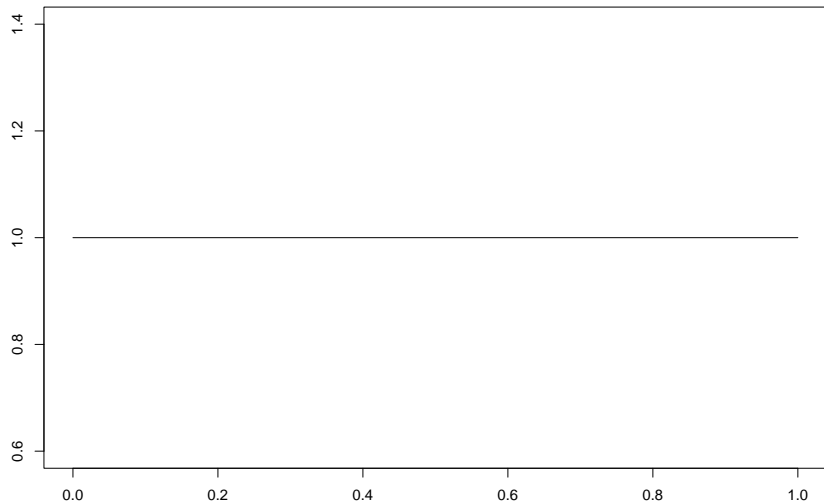
Skewed Right Distributions



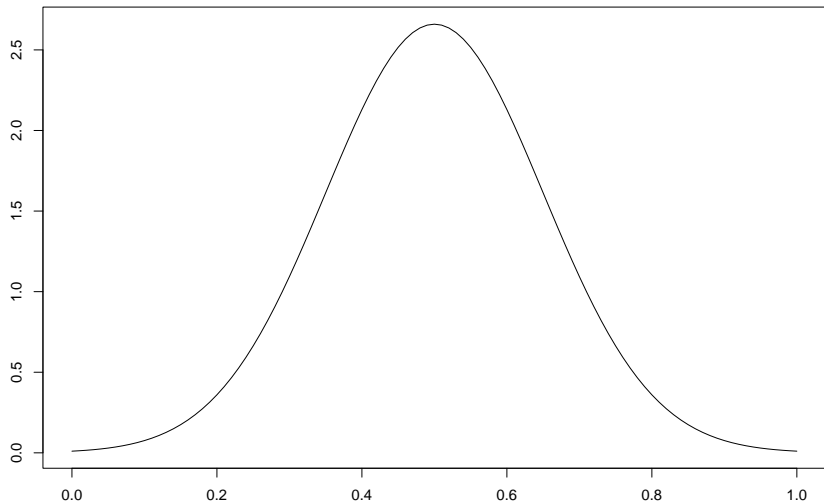
Skewed Left Distributions



Uniform Distributions

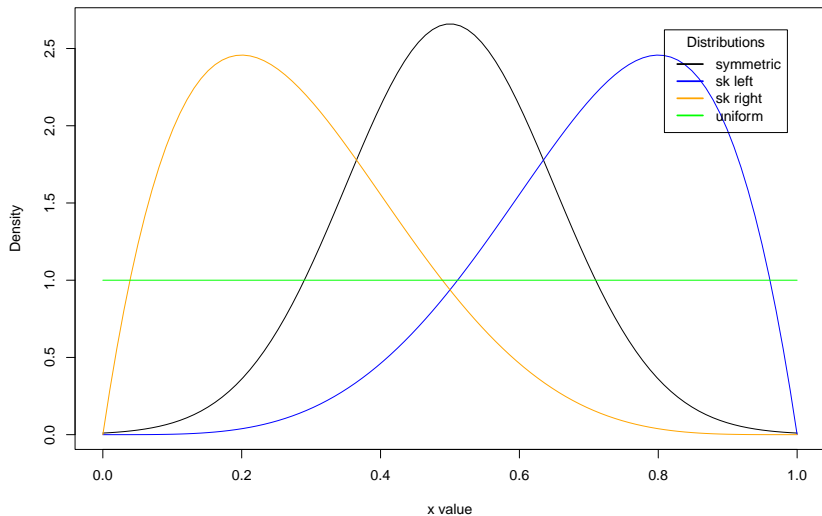


Symmetric and bell-shaped distributions



All together now!

Some density curves

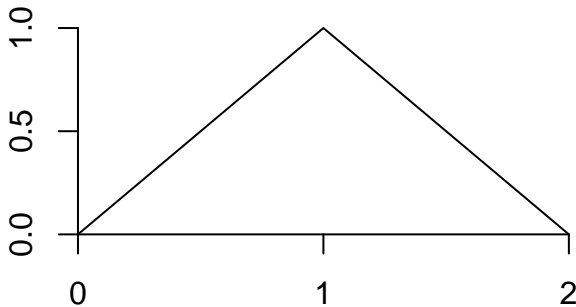


Continuous Random Variables

- ▶ A continuous random variable X has a density curve associated with it.
- ▶ $P(a \leq X \leq b) \leftrightarrow$ Area under density curve between a and b .
- ▶ Note: $P(X = a) = 0$, for any value a .

Example: A continuous random variable

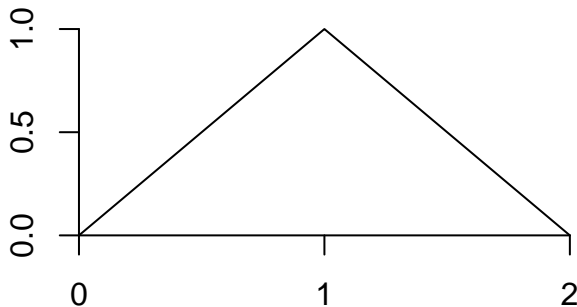
Suppose a random variable X has the following density curve



- ▶ Wait, is this a density curve?
 - ▶ Does it lie above x-axis?
 - ▶ Is the area underneath it equal to 1?

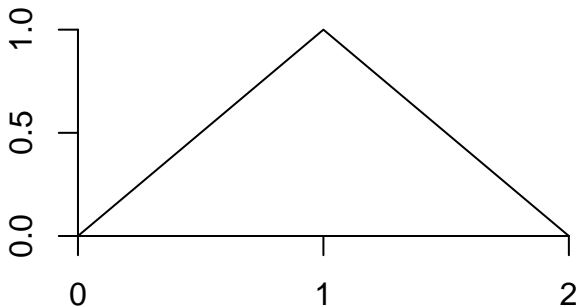
Example: A continuous random variable

Suppose a random variable X has the following density curve



- ▶ Wait, is this a density curve?
 - ▶ Does it lie above x-axis? [Check!](#)
 - ▶ Is the area underneath it equal to 1? [Check!](#)

Example: A continuous random variable

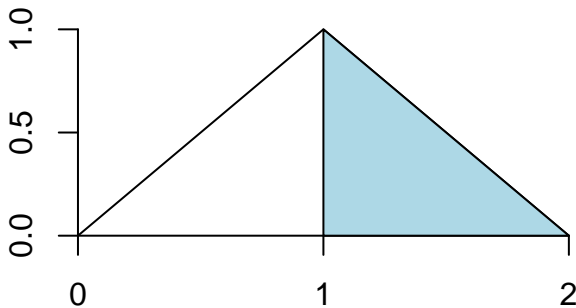


► Notes:

- The **values** of the distribution consist of all real numbers from 0 to 2.
- The **shape** tells me that values close to 1 are more likely to occur than values close to 0 or 2.

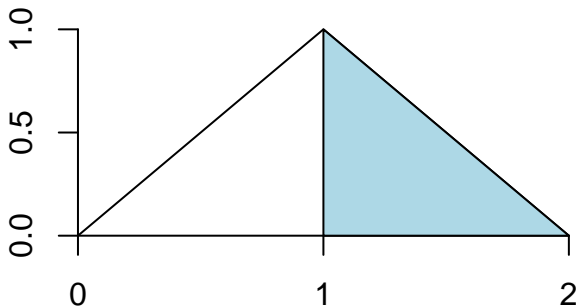
Area under a density curve corresponds to proportions

Question: What proportion of the distribution is between 1 and 2?
In other words, what is $P(1 \leq X \leq 2)$?



Area under a density curve corresponds to proportions

Question: What proportion of the distribution is between 1 and 2?
In other words, what is $P(1 \leq X \leq 2)$?



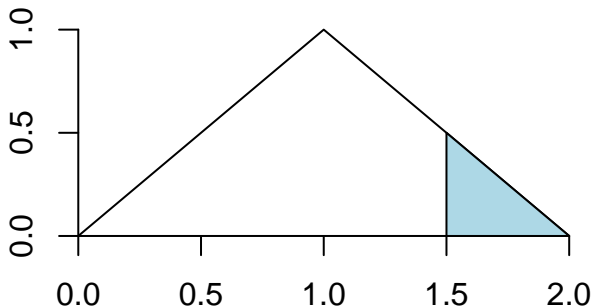
- ▶ This shaded region is a right triangle whose legs are both 1, so the area is

$$A = \frac{1}{2}(1)(1) = 0.5.$$

- ▶ So 50% of the distribution is between 1 and 2.
- ▶ We can also write $P(1 \leq X \leq 2) = 0.5$.

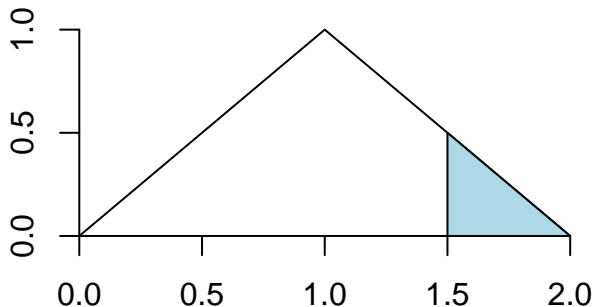
Area corresponds to proportion

Question: What proportion of the distribution has values between 1.5 and 2?



Area corresponds to proportion

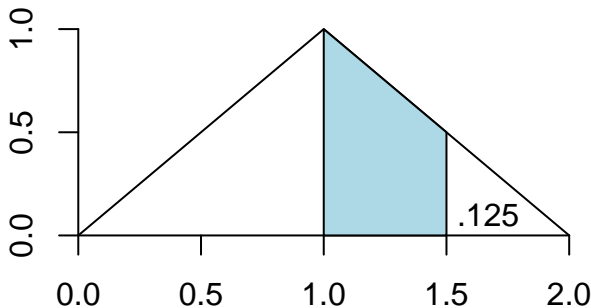
Question: What proportion of the distribution has values between 1.5 and 2?



- ▶ The shaded region above is also a right triangle, and $A = \frac{1}{2}(0.5)(0.5) = 0.125$.
- ▶ So 12.5% of the distribution is between 1.5 and 2.
- ▶ Put another way, $P(1.5 \leq X \leq 2) = 0.125$.

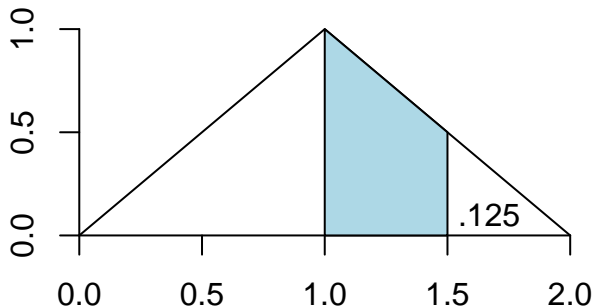
Area corresponds to proportion

Question: What proportion of the distribution has values between 1 and 1.5? What is $P(X = 1.5)$?



Area corresponds to proportion

Question: What proportion of the distribution has values between 1 and 1.5? What is $P(X = 1.5)$?



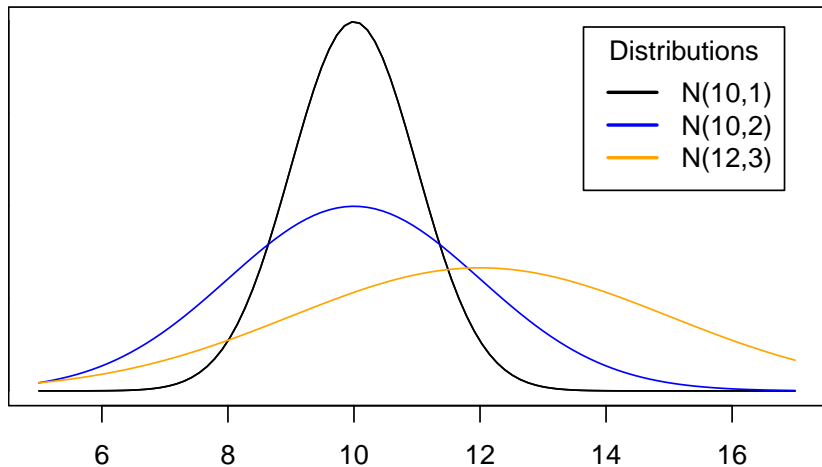
- ▶ By subtraction, $.5 - .125 = .375$, so $P(1 \leq X \leq 1.5) = 0.375$.
- ▶ $P(X = 1.5) = 0$ since for a continuous X , the probability of a particular value happening is always 0.

The Normal Distribution

The Normal Distribution

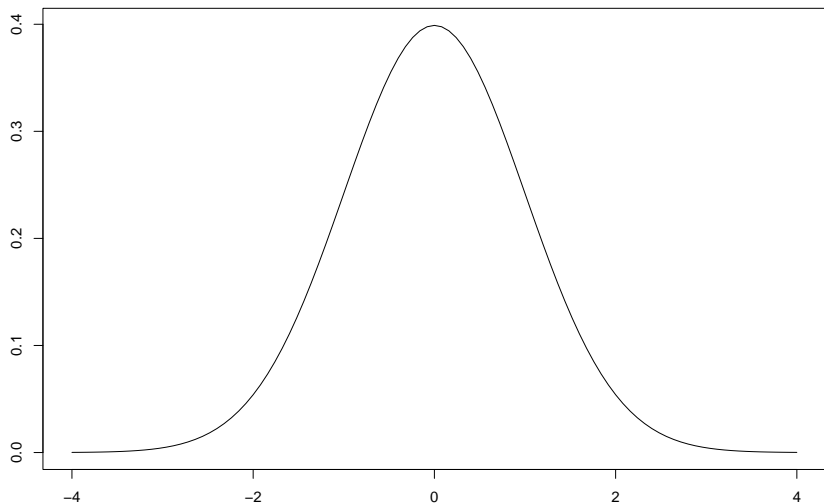
- ▶ A symmetric, bell-shaped distribution.
- ▶ Many variables are nearly normal, but none are exactly normal.
- ▶ We will use it in data exploration and to solve important problems in statistics.
- ▶ Its shape is uniquely determined by two parameters:
 - ▶ μ - mean
 - ▶ σ - standard deviation
- ▶ $N(\mu, \sigma)$ denotes this distribution

A few Normal Distributions



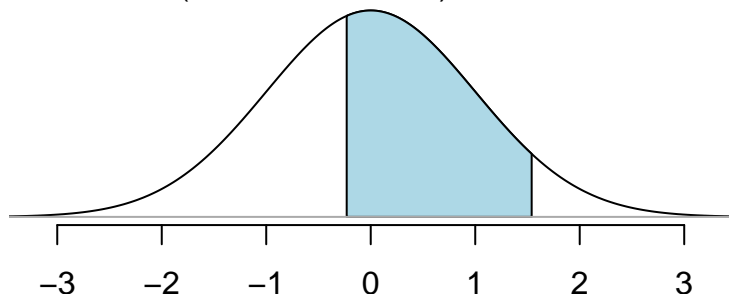
The standard normal distribution $N(0, 1)$

- ▶ Has mean $\mu = 0$, and standard deviation $\sigma = 1$
- ▶ It is of fundamental importance in statistics



Area under the standard normal distribution

$$P(-0.23 < z < 1.54) = 0.5292$$



- ▶ Let Z denote a value in $N(0, 1)$
- ▶ Let $P(-0.23 < Z < 1.54)$ denote the proportion of the distribution between -0.23 and 1.54.
- ▶ This region is not a simple geometric shape, but this area can be found using methods from Calculus.
- ▶ We have lots of resources available for computing areas under normal distributions, as we shall see.

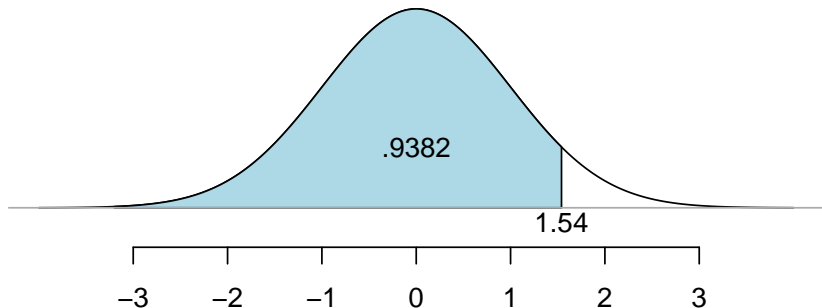
Computing areas under the standard normal distribution

Various Approaches

- ▶ Your graphing calculator (you can research this)
- ▶ RStudio (strongly recommend! I will demonstrate)
- ▶ Probability Table in book, p. 410-411 (old school)
- ▶ Other online app

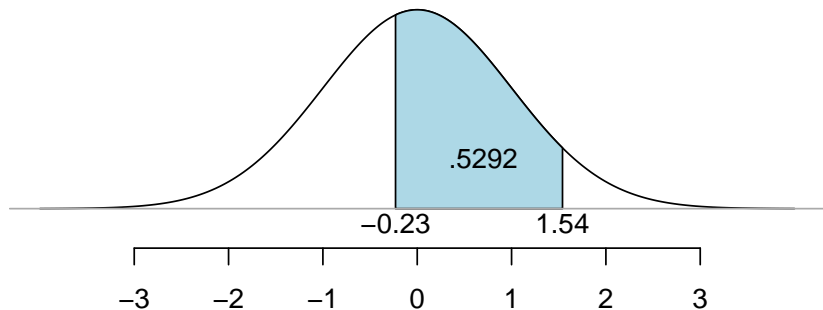
The `pnorm()` function in R

The `pnorm(z)` function gives the area to the left of z in $N(0,1)$



- ▶ `pnorm(1.54) = 0.9382`
- ▶ 93.82% of the distribution has a value less than $z = 1.54$.
- ▶ The **area to the right** of 1.54 is $1 - \text{pnorm}(1.54) = 0.0618$

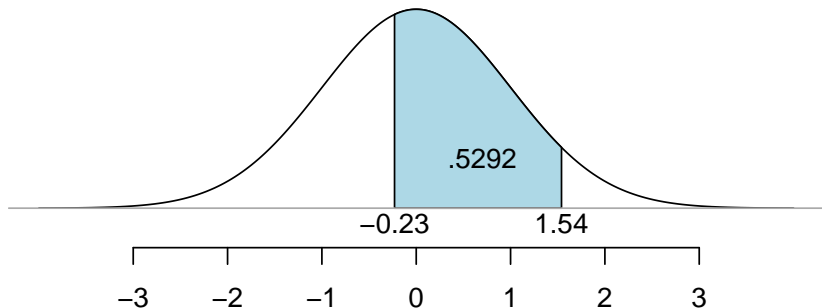
Find $P(-0.23 < Z < 1.54)$



```
pnorm(1.54)-pnorm(-0.23)
```

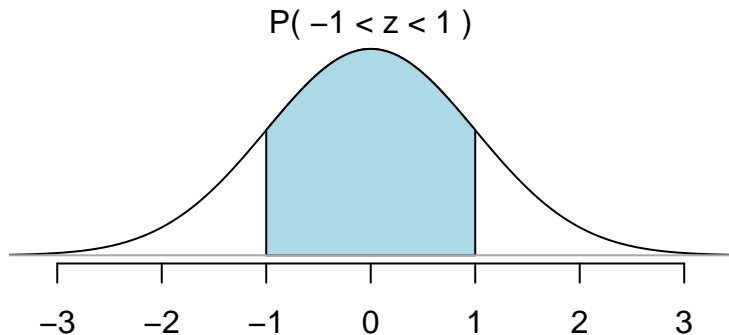
```
## [1] 0.5291739
```

Find $P(-0.23 < Z < 1.54)$



- ▶ Why does `pnorm(1.54)-pnorm(-0.23)` do the trick?
- ▶ Subtracting two “areas to the left” leaves the area in between.
- ▶ So 52.92% of the dist'n is between -0.23 and 1.54.

Example: Finding proportions in $N(0, 1)$.



We use R:

```
pnorm(1)-pnorm(-1)
```

```
## [1] 0.6826895
```

The 68-95-99.7 Rule

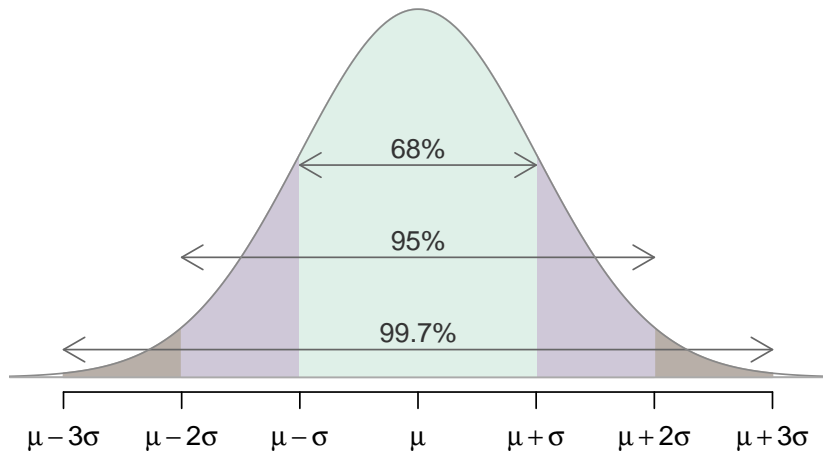
In any normal distribution $N(\mu, \sigma)$:

- ▶ About 68% of the distribution is within 1 standard deviation of the mean.
- ▶ About 95% of the distribution is within 2 standard deviations of the mean.
- ▶ About 99.7% of the distribution is within 3 standard deviations of the mean.

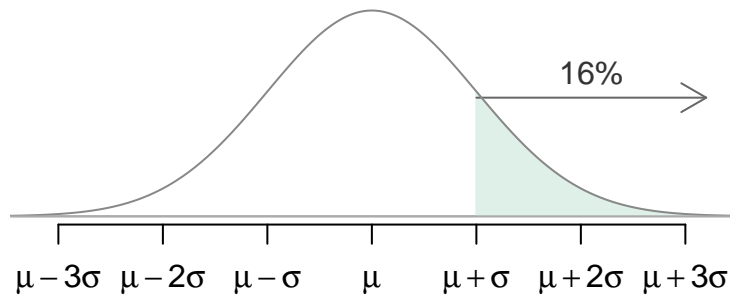
Example: In $N(10, 3)$,

- ▶ Roughly, 68% of the dist'n is between 7 and 13, and
- ▶ 95% of the dist'n is between 4 and 16, and
- ▶ 99.7% of the dist'n (almost all!) is between 1 and 19.

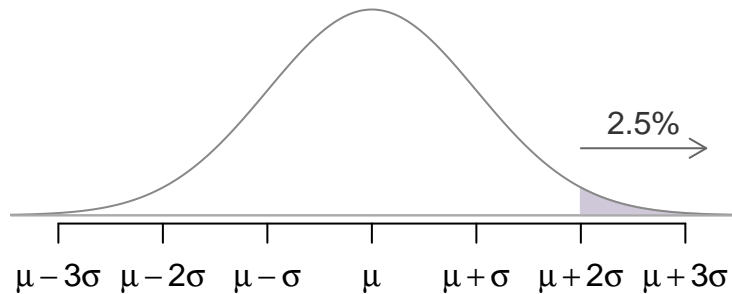
The 68-95-99.7 Rule



Upper Tail estimates



Upper Tail estimates



Estimating areas

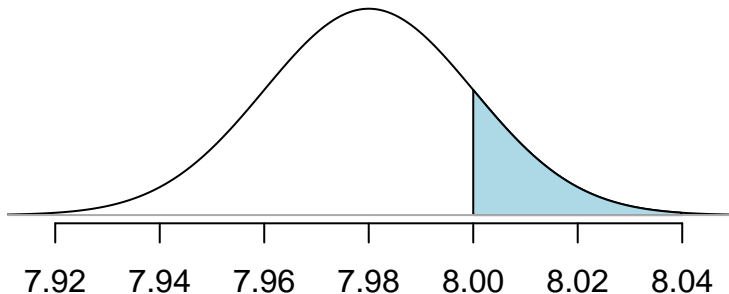
Example: Suppose the distribution of lengths of lumber labelled as 8 foot long 2x4s is normally distributed with $\mu = 7.98$ ft and $\sigma = 0.02$ ft.

1. The 68-95-99.7 rule tells us that essentially all boards sold will have length in a certain range. What is this range?
2. If I grab a board at random, what are the chances that it is actually 8 feet or longer?

Estimating areas

Solution

1. Nearly the entire distribution (99.7%) is within 3 standard deviations of the mean, meaning nearly every board has length in the range: $\mu - 3\sigma$ to $\mu + 3\sigma$. So, nearly every "8 ft 2 by 4" has length in the range from 7.92 ft to 8.04 ft.
2. 8 ft is exactly 1 st dev above the mean, so there is about a 16% chance that a random board has length at least 8 ft.

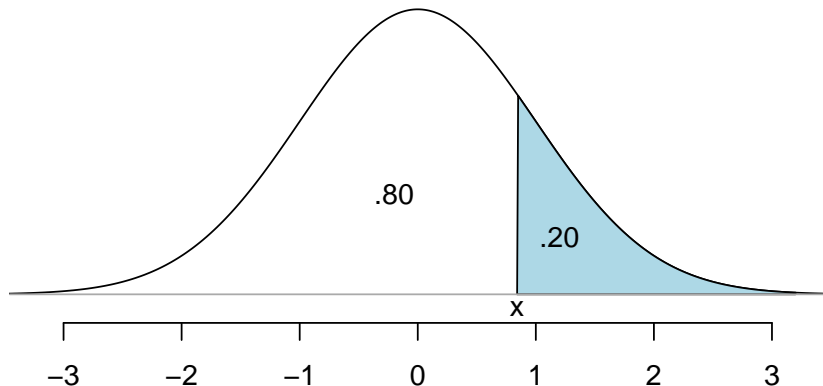


The `qnorm()` function in R.

- ▶ The `qnorm(A)` function in R gives the z -value in $N(0, 1)$ that has area A to the left of it.
- ▶ So it works “backwards” from the `pnorm()` function:
 - ▶ `pnorm(z)` returns the area A to the left of z
 - ▶ `qnorm(A)` returns the z -value that has area A to the left of it.
- ▶ For instance, what value of z is greater than 90% of the standard normal distribution?
- ▶ `qnorm(.9) = 1.28`
- ▶ So the area to the left of the z -value 1.28 equals 0.9.

Example with $qnorm(A)$

Find the value z in $N(0, 1)$ that has 20% of the distribution to the *right* of it.



Example with `qnorm(A)`

Solution 1

- ▶ Having 20% of the distribution to the right means having area .20 to the right, which means having area $1 - .20 = .80$ to the left.
- ▶ Compute `qnorm(.80)`, which is 0.8416212. So, the value $z \approx 0.84$ has about 20% of the distribution to the right of it.

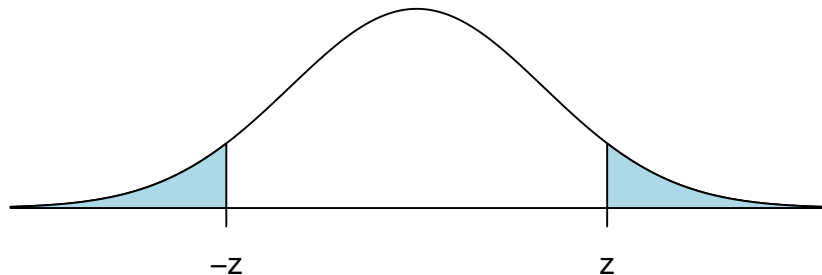
Solution 2

- ▶ Compute `-qnorm(.2)`, which is 0.8416212.

Why do these two approaches agree? Is it always true that $\text{qnorm}(1-A) = -\text{qnorm}(A)$?

The symmetry of the normal distribution

Useful fact For any value z , the area to the right of z equals the area to the left of $-z$. The two shaded areas are equal.



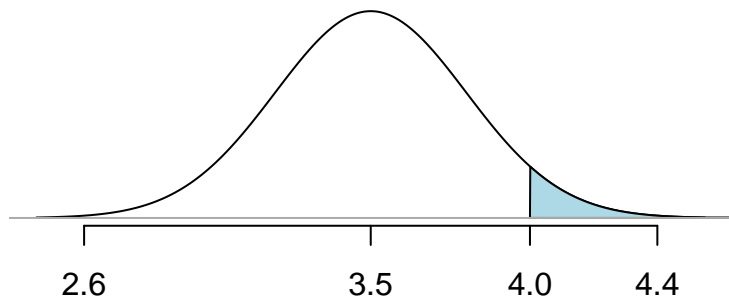
This tells us two things:

1. $\text{pnorm}(-z) = 1 - \text{pnorm}(z)$
2. $\text{qnorm}(1-A) = -\text{qnorm}(A)$

Other normal distributions $N(\mu, \sigma)$

Example Let X denote the wingspan of a monarch butterfly. Experience shows that the random variable X is approximately normal with mean $\mu = 3.5$ inches and standard deviation $\sigma = 0.3$ inches.

Q: What proportion of Monarch butterflies have wingspans greater than 4 inches? In other words, what is $P(X > 4.0)$?



Standardizing with Z-scores

If x is a value in $N(\mu, \sigma)$, its **Z-score** tells us how far away x is from μ in standard deviation units.

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

The Z-score of 4.0 in the butterfly wingspan distribution $N(3.5, 0.3)$ is

$$Z = \frac{4.0 - 3.5}{0.3} \approx 1.667.$$

This shows that in $N(3.5, 0.3)$ the value 4.0 is about 1.67 standard deviations away from the mean.

Standardizing with Z-scores

- ▶ We were asked to find the proportion of Monarch butterflies with wingspan greater than 4.0 inches.
- ▶ This proportion equals the area to the right of 4.0 in $N(3.5, 0.3)$.
- ▶ This area equals the area to the right of 1.667 in $N(0, 1)$.
- ▶ That is,

$$P(X > 4.0) = P(Z > 1.667),$$

and we use R to find this proportion:

```
1-pnorm(1.667)
```

```
## [1] 0.0477572
```

- ▶ About 4.8% of all Monarch butterflies have wingspan exceeding 4 inches.

The `pnorm()` function for $N(\mu, \sigma)$

- ▶ We can use `pnorm()` directly for any $N(\mu, \sigma)$, without first converting to Z -scores:

```
1-pnorm(4, mean=3.5, sd=0.3)
```

```
## [1] 0.04779035
```

- ▶ About 4.8% of all Monarch butterflies have wingspan exceeding 4 inches.
- ▶ Having said that, it is often helpful to compute Z -scores for values in $N(\mu, \sigma)$, and we will continue to do so.

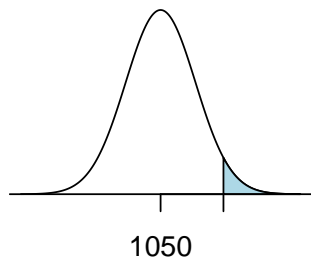
Example: Temperature in a kiln

- ▶ The temperature at any random location in a kiln used in the manufacture of bricks is normally distributed with mean 1050°F and standard deviation 50°F .
- ▶ That is, the temperature in the kiln follows the distribution $N(1050, 50)$.
- ▶ If bricks are fired at a temperature above 1140°F , they will crack and must be thrown away.
- ▶ If the bricks are placed randomly throughout the kiln, what proportion of bricks will crack during the firing process?

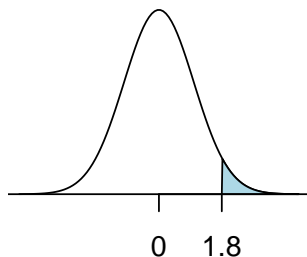
Example: Temperature in a kiln

- ▶ We want to find $P(X > 1140)$.
- ▶ First standardize with Z-scores:

$$P(X > 1140) = P\left(Z > \frac{1140 - 1050}{50}\right) = P(Z > 1.80).$$



$N(1050, 50)$



$N(0, 1)$

Example: Temperature in a kiln

- ▶ We use R to find this proportion:

```
1-pnorm(1.8)
```

```
## [1] 0.03593032
```

- ▶ About 3.6% of the bricks will crack during the firing process.
- ▶ Recall, the original question asked for the proportion of the $N(1050, 50)$ greater than 1140, and this can be found directly in R:

```
1-pnorm(1140, mean=1050, sd=50)
```

```
## [1] 0.03593032
```

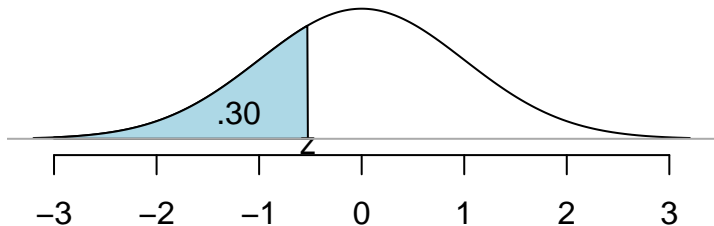
Finding benchmark values in $N(\mu, \sigma)$.

Example: Test scores on a remarkably fun stats exam are approximately normal with $\mu = 82$, $\sigma = 7$. How high does one need to score on the exam so that 30% of the class had a lower score?

We want to find the value L so that $P(X < L) = 0.30$.

Solution:

1. Find the Z-score in $N(0,1)$ that has area to the left = 0.30.
This is $z = \text{qnorm}(.3) = -0.5244005$



Finding benchmark values in $N(\mu, \sigma)$.

Solution:

2. Find the value of L in $N(82, 7)$ that has Z-score = -0.524. To do this, solve this equation for L :

$$\frac{L - 82}{7} = -0.524.$$

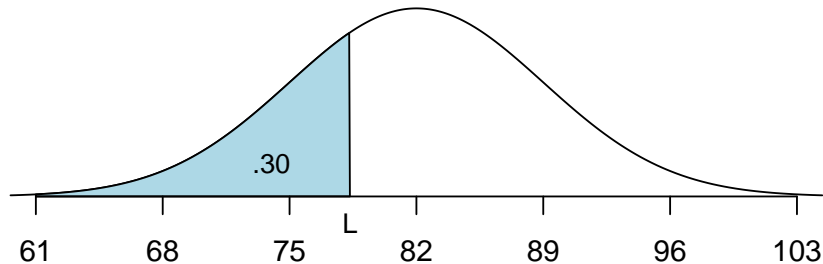
We multiply both sides by 7, then add 82 to both sides, to obtain:

$$L = 7 \cdot (-0.524) + 82 \approx 78.3.$$

Finding benchmark values in $N(\mu, \sigma)$.

Solution:

3. We have found that in $N(82, 7)$, 30% of the distribution will be less than $L = 78.3$.



4. In other words, *30% of the scores on this remarkably fun stats test will be less than 78.3.*