The Standard Deviation as a Ruler and the Normal Model

Al Nosedal.
University of Toronto.

Summer 2019

My momma always said: "Life was like a box of chocolates. You never know what you're gonna get."

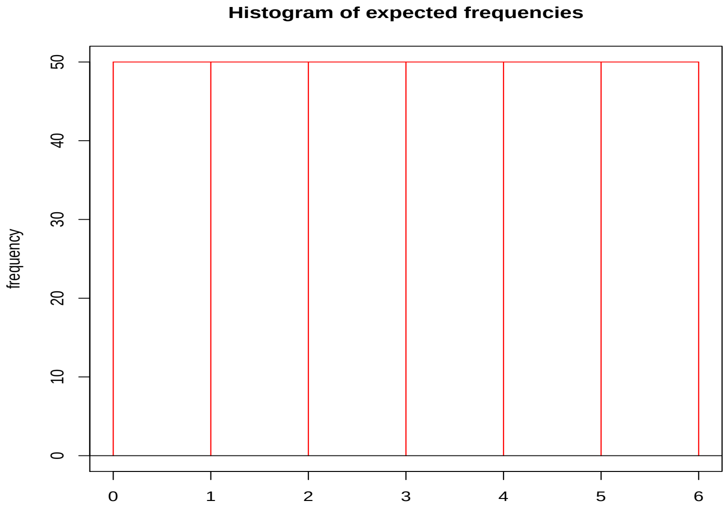
Forrest Gump.

Simple Example

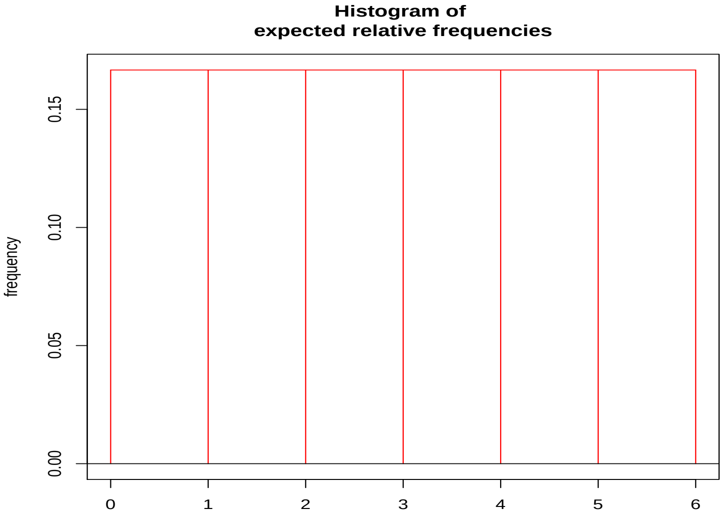
Random Experiment: Rolling a fair die 300 times.

Class	Expected Frequency	Expected Relative Freq
$0 < x \leq 1$	50	$1/6$
$1 < x \leq 2$	50	$1/6$
$2 < x \leq 3$	50	$1/6$
$3 < x \leq 4$	50	$1/6$
$4 < x \leq 5$	50	$1/6$
$5 < x \leq 6$	50	$1/6$

Histogram of Expected Frequencies



Histogram of Expected Relative Frequencies

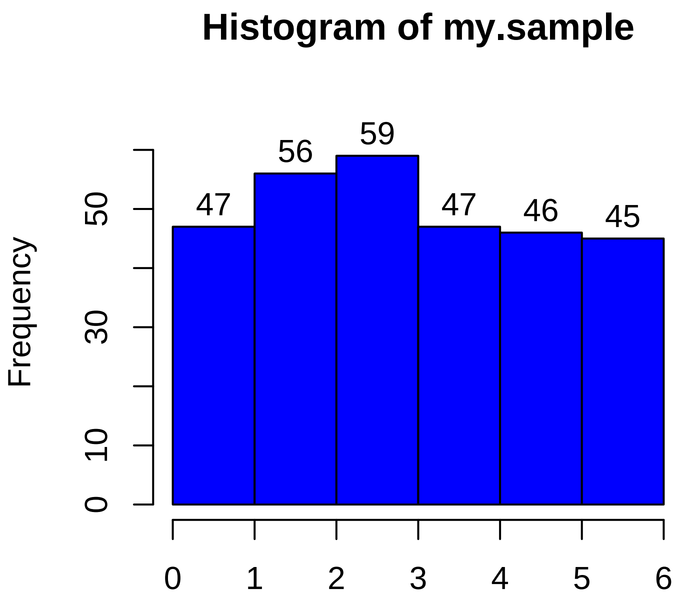


Six simulations

```
die=c(1,2,3,4,5,6);
sample(die,1,replace=TRUE);

## [1] 2

sample(die,6,replace=TRUE);

## [1] 4 3 1 2 1 1
```

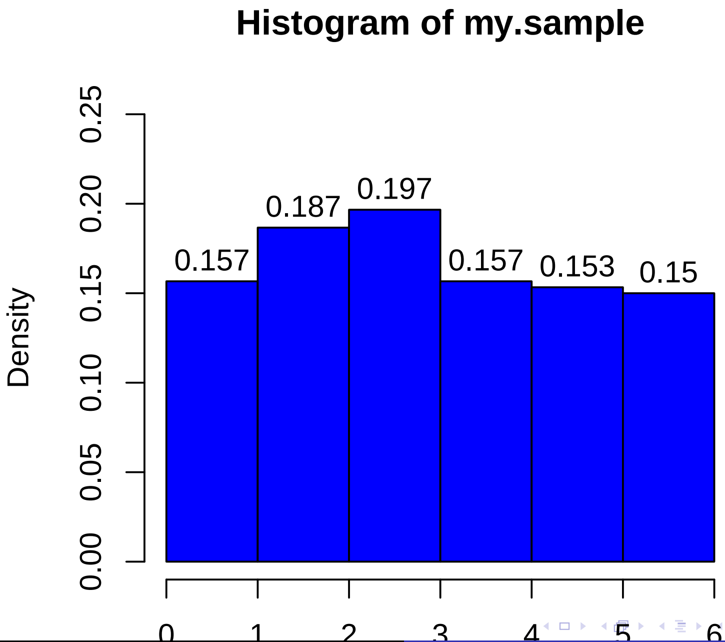
Histogram (frequencies)

```
die=c(1,2,3,4,5,6);
my.sample=sample(die,300,replace=TRUE);
classes=seq(0,6,by=1);
hist(my.sample,breaks=classes,labels=TRUE,
col="blue",ylim=c(0,65));
```


Histogram of my.sample
Frequency
0
10
30
50
47
56
59
47
46
45
0
1
2
3
4
5
6

Histogram (relative frequencies)

```
classes=seq(0,6,by=1);
hist(my.sample,breaks=classes,labels=TRUE,
col="blue",freq=FALSE,ylim=c(0,0.25));
```


Histogram of my.sample
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.157
0.187
0.197
0.157
0.153
0.15
0
1
2
3
4
5
6

Density Curve

A density curve is a curve that is always on or above the horizontal axis, and has area exactly 1 underneath it.
A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range.
Note. No set of real data is exactly described by a density curve. The curve is an idealized description that is easy to use and accurate enough for practical use.

Examining the location of accidents on a level, 5-mile bike path shows that they occur uniformly along the length of the path. The figure below displays the density curve that describes the distribution of accidents.

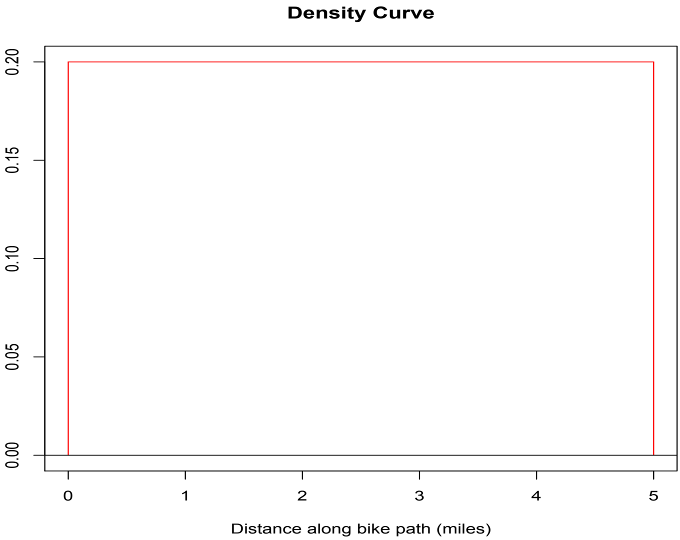
a) Explain why this curve satisfies the two requirements for a density curve.

b) The proportion of accidents that occur in the first mile of the path is the area under the density curve between 0 miles and 1 mile. What is this area?

c) There is a stream alongside the bike path between the 0.8-mile mark and the 1.3-mile mark. What proportion of accidents happen on the bike path alongside the stream?

d) The bike path is a paved path through the woods, and there is a road at each end. What proportion of accidents happen more than 1 mile from either road?

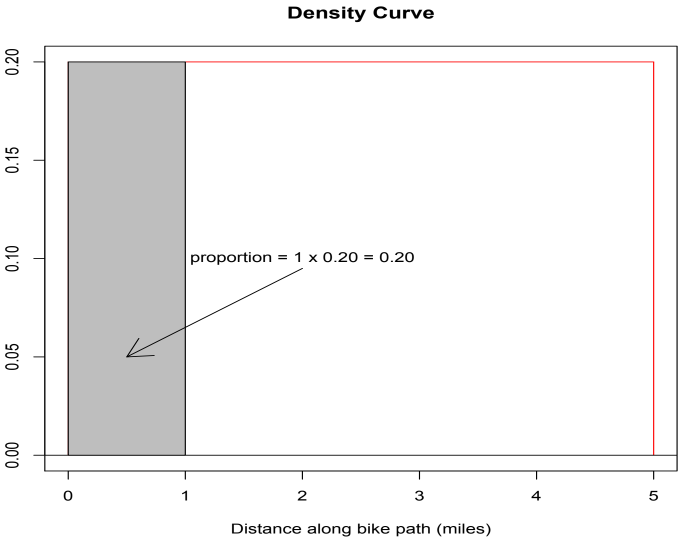
Density Curve



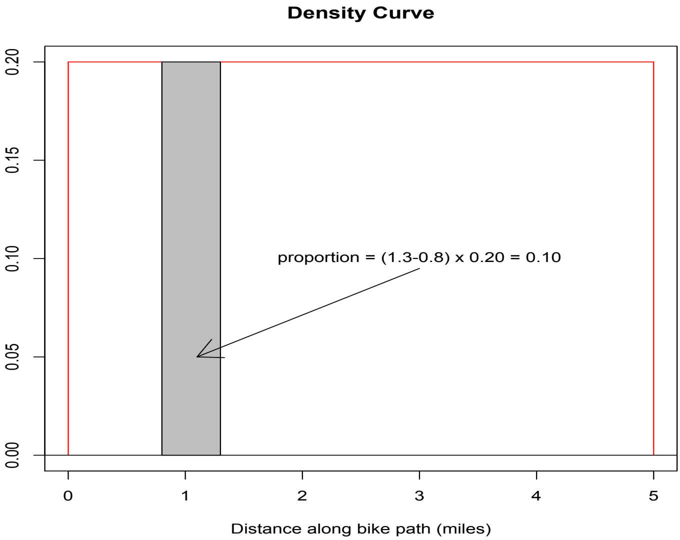
Solution

a) It is on or above the horizontal axis everywhere, and because it forms a $1/5 \times 5$ rectangle, the area beneath the curve is 1.

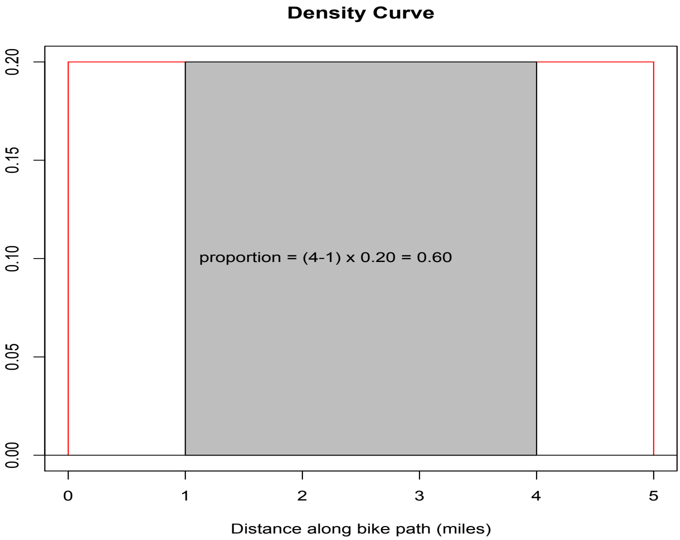
Solution b)


Density Curve
0.00
0.05
0.10
0.15
0.20
proportion = 1 x 0.20 = 0.20
0
1
2
3
4
5
Distance along bike path (miles)

Solution c)


Density Curve
0.00
0.05
0.10
0.15
0.20
proportion = (1.3-0.8) x 0.20 = 0.10
0
1
2
3
4
5
Distance along bike path (miles)

Solution d)


Density Curve
proportion = (4-1) x 0.20 = 0.60
0.00
0.05
0.10
0.15
0.20
0
1
2
3
4
5
Distance along bike path (miles)

HW?

The amount of gasoline sold daily at a service station is uniformly distributed with a minimum of 2,000 gallons and a maximum of 5,000 gallons.

a. Find the probability that daily sales will fall between 2,500 and 3,000 gallons.

b. What is the probability that the service station will sell at least 4,000 gallons?

c. What is the probability that the station will sell exactly 2,500 gallons?

Solution

a. $P(2500 \leq X \leq 3000) = (3000 - 2500)\left(\frac{1}{3000}\right) = 0.1667.$

b. $P(X \geq 4000) = (5000 - 4000)\left(\frac{1}{3000}\right) = 0.3333.$

c. $P(X = 2500) = 0.$

Normal Density Function

The probability density function of a **Normal random variable** is

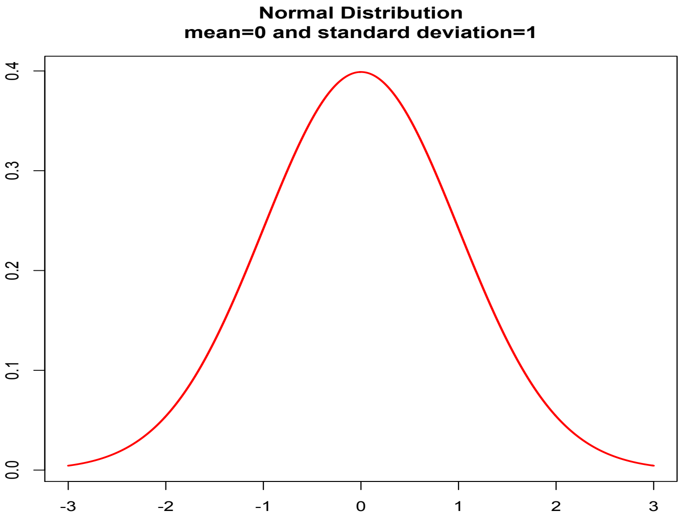
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $-\infty < x < \infty$, $e = 2.71828...$ and $\pi = 3.14159...$

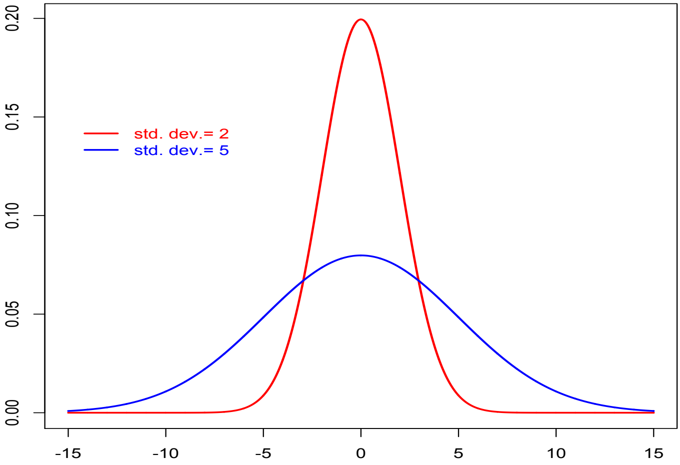
Normal Distributions

A Normal Distribution is described by a Normal density curve. Any particular Normal distribution is completely specified by two numbers, its mean μ and standard deviation σ.
The mean of a Normal distribution is at the center of the symmetric Normal curve. The standard deviation is the distance from the center to the change-of-curvature points on either side.

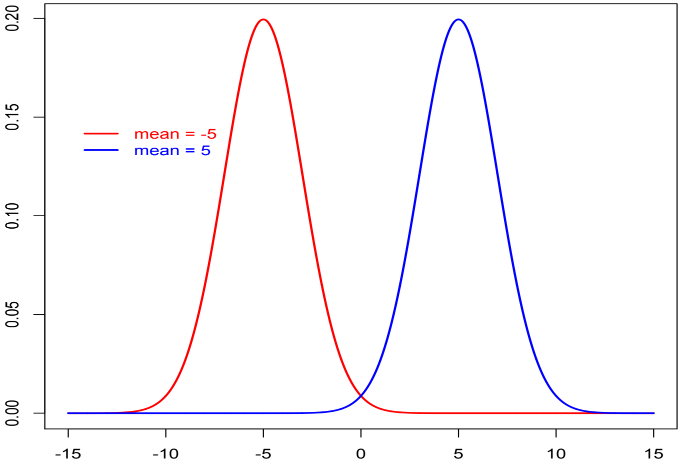
Standard Normal Distribution



Two Different Standard Deviations



Two Different Means



The 68-95-99.7 rule

In a Normal distribution with mean μ and standard deviation σ:
Approximately 68% of the observations fall within σ of the mean μ.
Approximately 95% of the observations fall within 2σ of μ.
Approximately 99.7% of the observations fall within 3σ of μ.

The national average for the verbal portion of the College Boards Scholastic Aptitude Test (SAT) is 507. The College Board periodically rescales the test scores such that the standard deviation is approximately 100. Answer the following questions using a bell-shaped distribution and the empirical rule for the verbal test scores.

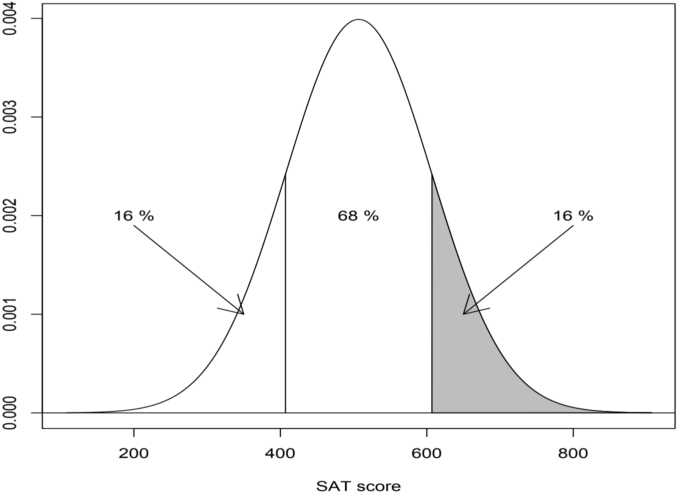
a. What percentage of students have an SAT verbal score greater than 607?

b. What percentage of students have an SAT verbal score greater than 707?

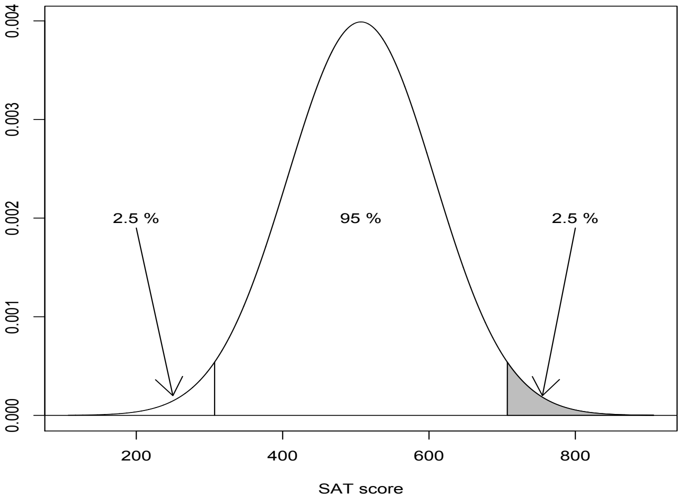
c. What percentage of students have an SAT verbal score between 407 and 507?

d. What percentage of students have an SAT verbal score between 307 and 707?

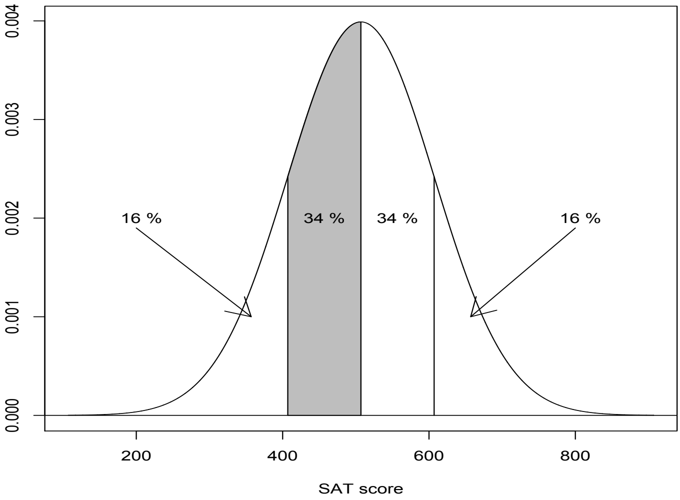
Solution a)



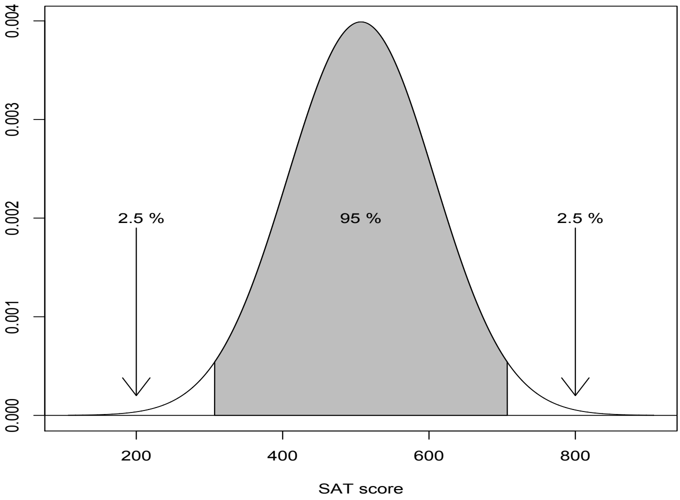
Solution b)



Solution c)



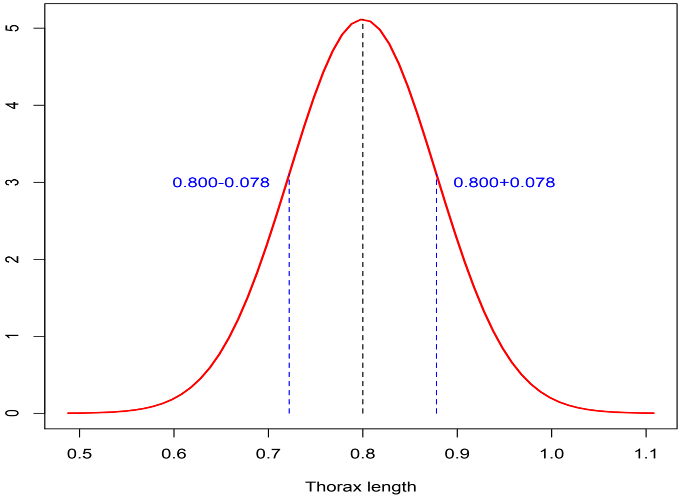
Solution d)



Fruit flies

The common fruit fly Drosophila melanogaster is the most studied organism in genetic research because it is small, easy to grow, and reproduces rapidly. The length of the thorax (where the wings and legs attach) in a population of male fruit flies is approximately Normal with mean 0.800 millimeters (mm) and standard deviation 0.078 mm. Draw a Normal curve on which this mean and standard deviation are correctly located.

Solution



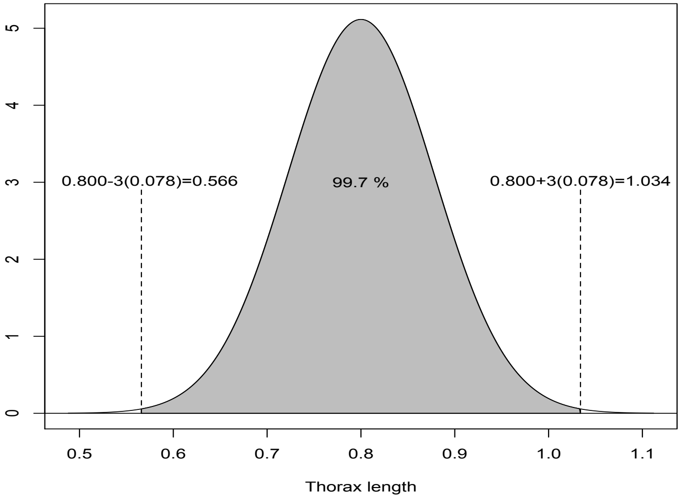
Fruit flies

The lenght of the thorax in a population of male fruit flies is approximately Normal with mean 0.800 mm and standard deviation 0.078 mm. Use the 68-95-99.7 rule to answer the following questions.

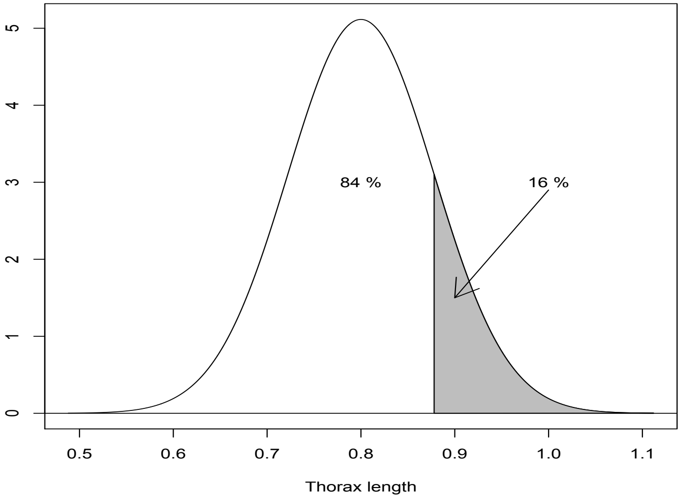
a) What range of lengths covers almost all (99.7%) of this distribution?

b) What percent of male fruit flies have a thorax length exceeding 0.878 mm?

Solution a) Between 0.566 mm and 1.034 mm



Solution b) 16% of thorax lenghts exceed 0.878 mm



Monsoon rains (HW?)

The summer monsoon brings 80% of India's rainfall and is essential for the country's agriculture. Records going back more than a century show that the amount of monsoon rainfall varies from the year according to a distribution that is approximately Normal with mean 582 mm and standard deviation 82 mm. Use the 68-95-99.7 rule to answer the following questions.

a) Between what values do the monsoon rains fall in 95% of all years?

b) How small are the monsoon rains in the dryest 2.5% of all years?

Solution

a) In 95% of all years, monsoon rain levels are between
582 - 2(82) and 582 + 2(82) i.e. 418 mm and 746 mm.
b) The driest 2.5% of monsoon rainfalls are less than 418 mm; this is more than two standard deviations below the mean.

Standard Normal Distribution

The standard Normal distribution is the Normal distribution $N(0, 1)$ with mean 0 and standard deviation 1.
If a variable x has any Normal distribution N(μ,σ) with mean μ and standard deviation σ, then the standardized variable

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

has the standard Normal distribution.

SAT vs ACT

In 2010, when she was a high school senior, Alysha scored 670 on the Mathematics part of the SAT. The distribution of SAT Math scores in 2010 was Normal with mean 516 and standard deviation 116. John took the ACT and scored 26 on the Mathematics portion. ACT Math scores for 2010 were Normally distributed with mean 21.0 and standard deviation 5.3. Find the standardized scores for both students. Assuming that both tests measure the same kind of ability, who had the higher score?

Solution

Alysha's standardized score is

$$z_A = \frac{670 - 516}{116} = 1.33.$$

John's standardized score is

$$z_J = \frac{26 - 21}{5.3} = 0.94.$$

Alysha's score is relatively higher than John's.

Men's and women's heights

The heights of women aged 20 to 29 are approximately Normal with mean 64.3 inches and standard deviation 2.7 inches. Men the same age have mean height 69.9 inches with standard deviation 3.1 inches. What are the z-scores for a woman 6 feet tall and a man 6 feet tall? Say in simple language what information the z-scores give that the original nonstandardized heights do not.

Solution

We need to use the same scale, so recall that 6 feet = 72 inches. A woman 6 feet tall has standardized score

$$z_W = \frac{72 - 64.3}{2.7} = 2.85$$

(quite tall, relatively).
A man 6 feet tall has standardized score

$$z_M = \frac{72 - 69.9}{3.1} = 0.68.$$

Hence, a woman 6 feet tall is 2.85 standard deviations taller than average for women. A man 6 feet tall is only 0.68 standard deviations above average for men.

Use table 3 to find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

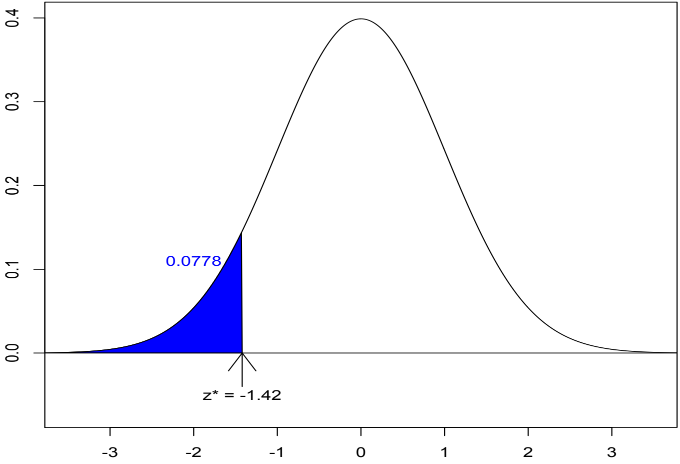
a) $z < -1.42$

b) $z > -1.42$

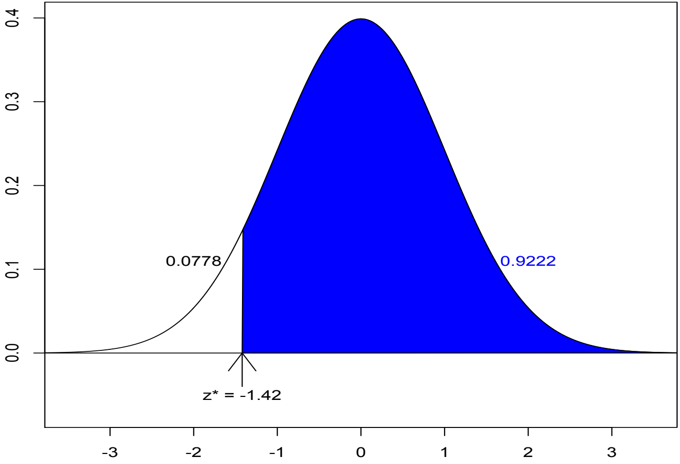
c) $z < 2.35$

d) $-1.42 < z < 2.35$

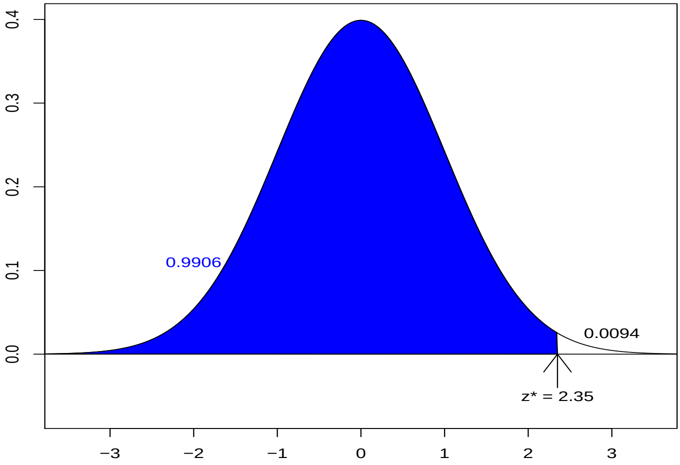
Solution a) 0.0778



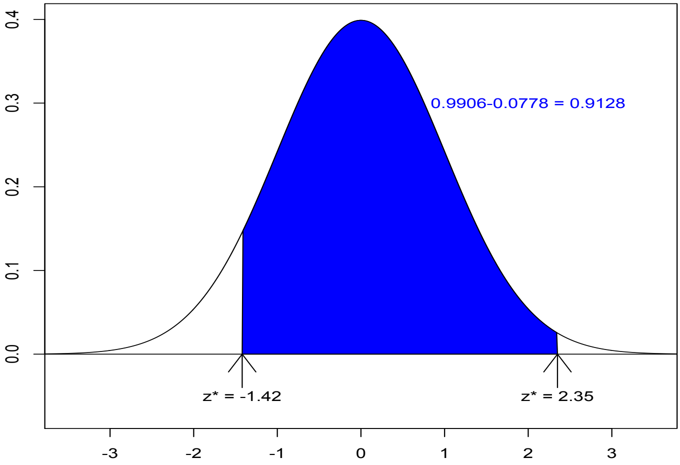
Solution b) 0.9222



Solution c) 0.9906



Solution d) 0.9966 - 0.0778 = 0.9128



Monsoon rains

The summer monsoon rains in India follow approximately a Normal distribution with mean 852 mm of rainfall and standard deviation 82 mm.

a) In the drought year 1987, 697 mm of rain fell. In what percent of all years will India have 697 mm or less of monsoon rain?

b) "Normal rainfall" means within 20% of the long-term average, or between 683 and 1022 mm. In what percent of all years is the rainfall normal?

Solution a)

1. State the problem. Let x be the monsoon rainfall in a given year. The variable x has the $N(852, 82)$ distribution. We want the proportion of years with $x \leq 697$.
2. Standardize. Subtract the mean, then divide by the standard deviation, to turn x into a standard Normal z.
Hence $x \leq 697$ corresponds to $z \leq \frac{697-852}{82} = -1.89$.
3. Use the table. From Table A, we see that the proportion of observations less than -1.89 is 0.0294. Thus, the answer is 2.94%.

Solution b)

1. State the problem. Let x be the monsoon rainfall in a given year. The variable x has the $N(852, 82)$ distribution. We want the proportion of years with $683 < x < 1022$.
2. Standardize. Subtract the mean, then divide by the standard deviation, to turn x into a standard Normal z.
$683 < x < 1022$ corresponds to $\frac{683-852}{82} < z < \frac{1022-852}{82}$, or $-2.06 < z < 2.07$.
3. Use the table. Hence, using Table A, the area is $0.9808 - 0.0197 = 96.11\%$.

The Medical College Admission Test

Almost all medical schools in the United States require students to take the Medical College Admission Test (MCAT). The exam is composed of three multiple-choice sections (Physical Sciences, Verbal Reasoning, and Biological Sciences). The score on each section is converted to a 15-point scale so that the total score has a maximum value of 45. The total scores follow a Normal distribution, and in 2010 the mean was 25.0 with a standard deviation of 6.4. There is little change in the distribution of scores from year to year.

a) What proportion of students taking the MCAT had a score over 30?

b) What proportion had scores between 20 and 25?

Solution a)

1. State the problem. Let x be the MCAT score of a randomly selected student. The variable x has the $N(25, 6.4)$ distribution. We want the proportion of students with $x > 30$.
2. Standardize. Subtract the mean, then divide by the standard deviation, to turn x into a standard Normal z.
Hence $x > 30$ corresponds to $z > \frac{30-25}{6.4} = 0.78$.
3. Use the table. From Table A, we see that the proportion of observations less than 0.78 is 0.7823. Hence, the answer is $1 - 0.7823 = 0.2177$, or 21.77%.

Solution b)

1. State the problem. Let x be the MCAT score of a randomly selected student. The variable x has the $N(25, 6.4)$ distribution. We want the proportion of students with $20 \leq x \leq 25$.
2. Standardize. Subtract the mean, then divide by the standard deviation, to turn x into a standard Normal z.
$20 \leq x \leq 25$ corresponds to $\frac{20-25}{6.4} \leq z \leq \frac{25-25}{6.4}$, or $-0.78 \leq z \leq 0$.
3. Use the table. Using Table A, the area is
$0.5 - 0.2177 = 0.2833$, or 28.33%.

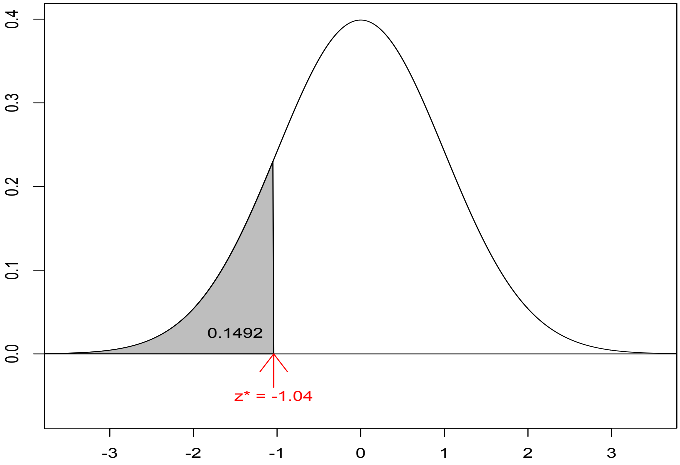
Step 1. State the problem in terms of the observed variable x. Draw a picture that shows the proportion you want in terms of cumulative proportions.
Step 2. Standardize x to restate the problem in terms of a standard Normal variable z.
Step 3. Use Table A and the fact that the total are under the curve is 1 to find the required area under the standard Normal curve.

Use Table A to find the value $z*$ of a standard Normal variable that satisfies each of the following conditions. (Use the value of $z*$ from Table A that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of $z*$ marked on the axis.

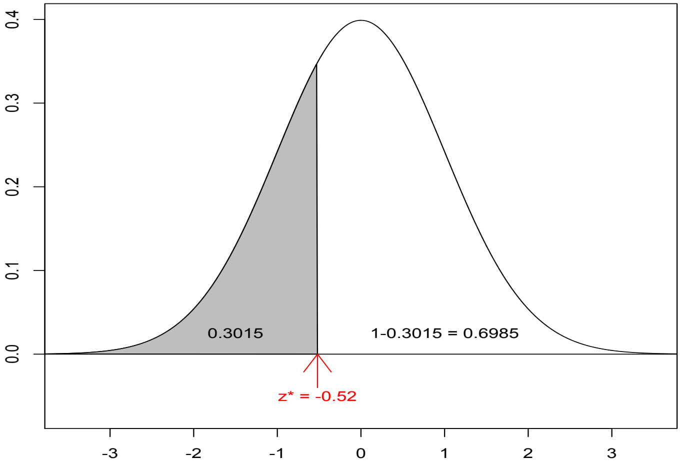
a) The point $z*$ with 15% of the observations falling below it.

b) The point $z*$ with with 70% of the observations falling above it.

Solution a) $z^* = -1.04$



Solution b) $z^* = -0.52$



The total scores on the Medical College Admission Test (MCAT) follow a Normal distribution with mean 25.0 and standard deviation 6.4. What are the median and the first and third quartiles of the MCAT scores?

Solution: Finding the median

Because the Normal distribution is symmetric, its median and mean are the same. Hence, the median MCAT score is 25.

Solution: Finding Q_1

1. State the problem. We want to find the MCAT score x with area 0.25 to its left under the Normal curve with mean $\mu = 25$ and standard deviation $\sigma = 6.4$.
2. Use the table. Look in the body of Table A for the entry closest to 0.25. It is 0.2514. This is the entry corresponding to $z* = -0.67$. So $z* = -0.67$ is the standardized value with area 0.25 to its left.
3. Unstandardize to transform the solution from the $z*$ back to the original x scale. We know that the standardized value of the unknown x is $z* = -0.67$.
So x itself satisfies

$$\frac{x - 25}{6.4} = -0.67$$

Solving this equation for x gives
$x = 25 + (-0.67)(6.4) = 20.71$

Solution: Finding Q_3

1. State the problem. We want to find the MCAT score x with area 0.75 to its left under the Normal curve with mean $\mu = 25$ and standard deviation $\sigma = 6.4$.
2. Use the table. Look in the body of Table A for the entry closest to 0.75. It is 0.7486. This is the entry corresponding to $z* = 0.67$. So $z* = 0.67$ is the standardized value with area 0.75 to its left.
3. Unstandardize to transform the solution from the $z*$ back to the original x scale. We know that the standardized value of the unknown x is $z* = 0.67$.
So x itself satisfies

$$\frac{x - 25}{6.4} = 0.67$$

Solving this equation for x gives
$x = 25 + (0.67)(6.4) = 29.29$

Finding a value when given a proportion

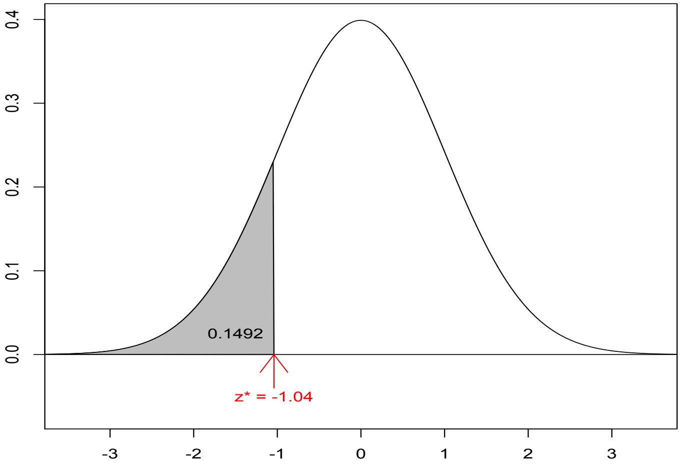
1. State the problem.
2. Use the table.
3. Unstandardize to transform the solution from the $z*$ back to the original x scale.

Use Table 3 to find the value $z*$ of a standard Normal variable that satisfies each of the following conditions. (Use the value of $z*$ from Table 3 that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of $z*$ marked on the axis.

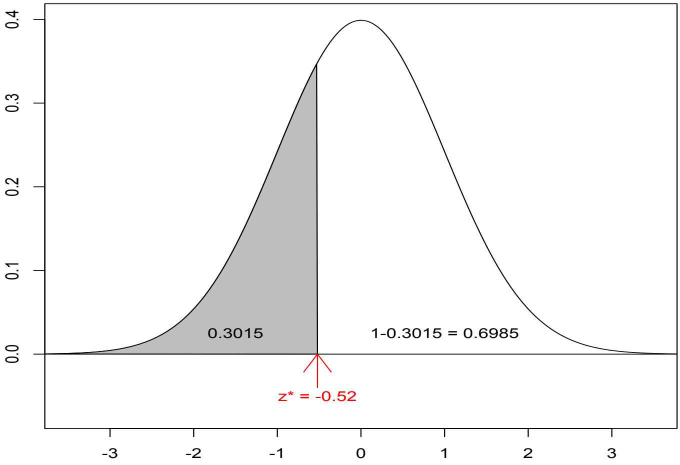
a) The point $z*$ with 15% of the observations falling below it.

b) The point $z*$ with with 70% of the observations falling above it.

Solution a) $z^* = -1.04$



Solution b) z* = -0.52



Example

The summer monsoon rains in India follow approximately a **Normal** distribution with mean 852 mm of rainfall and standard deviation 82 mm. a) In the drought year 1987, 697 mm of rain fell. In what percent of all years will India have 697 mm or less of monsoon rain? b) "Normal rainfall" means within 20% of the long-term average, or between 683 and 1022 mm. In what percent of all years is the rainfall normal?

R code

Just type the following:

```
# a)
pnorm(697, mean = 852, sd = 82);

## [1] 0.02936267

# b)
pnorm(1022, 852, 82) - pnorm(683, 852, 82);

## [1] 0.9612691
```

Example

Mensa is an organization whose members possess IQs that are in the top 2% of the population. It is known that IQs are Normally distributed with a mean of 100 and a standard deviation of 16. Find the minimum IQ needed to be a Mensa member.

Solution

Y = IQ. We know that Y has a Normal distribution with mean 100 and standard deviation 16.

1. State the problem. We want to find an IQ, y^*, with area 0.98 to its **left** under the Normal curve with mean $\mu = 100$ and standard deviation $\sigma = 16$.

2. Use the table. Look in the body of Table A for the entry closest to 0.98. It is .9798. This is the entry corresponding to $z* = 2.05$. So $z* = 2.05$ is the standardized value with area 0.98 to its left.

3. Unstandardize to transform the solution from the $z*$ back to the original Y scale. We know that the standardized value of the unknown y^* is $z* = 2.05$.
So y^* itself satisfies

$$\frac{y^* - 100}{16} = 2.05$$

Solving this equation for x gives
$y^* = 100 + (2.05)(16) = 132.8$

Solution (Formula)

$IQ^* = \mu + z^*\sigma = 100 + 2.05(16) = 132.8$
We could round it to 133.

R Code

```
qnorm(0.98, mean = 100, sd = 16);

## [1] 132.86
```

Cholesterol (HW?)

Assume the cholesterol levels of adult women can be described by a Normal model with a mean of 188 mg/dL and a standard deviation of 24.

a) What percent of adult women do you expect to have cholesterol levels over 200 mg/dL?

b) What percent of adult women do you expect to have cholesterol levels between 150 and 170 mg/dL?

c) Estimate the interquartile range of the cholesterol levels.

d) Above what value are the highest 15% of women's cholesterol levels?

Another example

Consider an investment whose return is Normally distributed with a mean of 10% and a standard deviation of 5%.
a. Determine the probability of losing money.
b. Find the probability of losing money when the standard deviation is equal to 10%.

Solution a)

X= return. We know that X is Normally distributed with mean (μ) 0.10 and standard deviation (σ) 0.05. Another way of writing this is: X has a $N(0.10, 0.05)$ distribution.

a. $P(X < 0) = P\left(\frac{X-\mu}{\sigma} < \frac{0-0.1}{0.05}\right) = P(Z < -2) = 0.0228.$

Therefore, the probability of losing money is 2.28%.

(Using the Empirical Rule gives 2.5%)

Solution b)

(If we increase the standard deviation to 10%).
$Y =$ return. We know that Y is Normally distributed with mean (μ) 0.10 and standard deviation (σ) 0.10. Another way of writing this is: Y has a $N(0.10, 0.10)$ distribution.

a. $P(Y < 0) = P\left(\frac{Y-\mu}{\sigma} < \frac{0-0.1}{0.1}\right) = P(Z < -1) = 0.1587$.

Therefore, the probability of losing money is 15.87%.