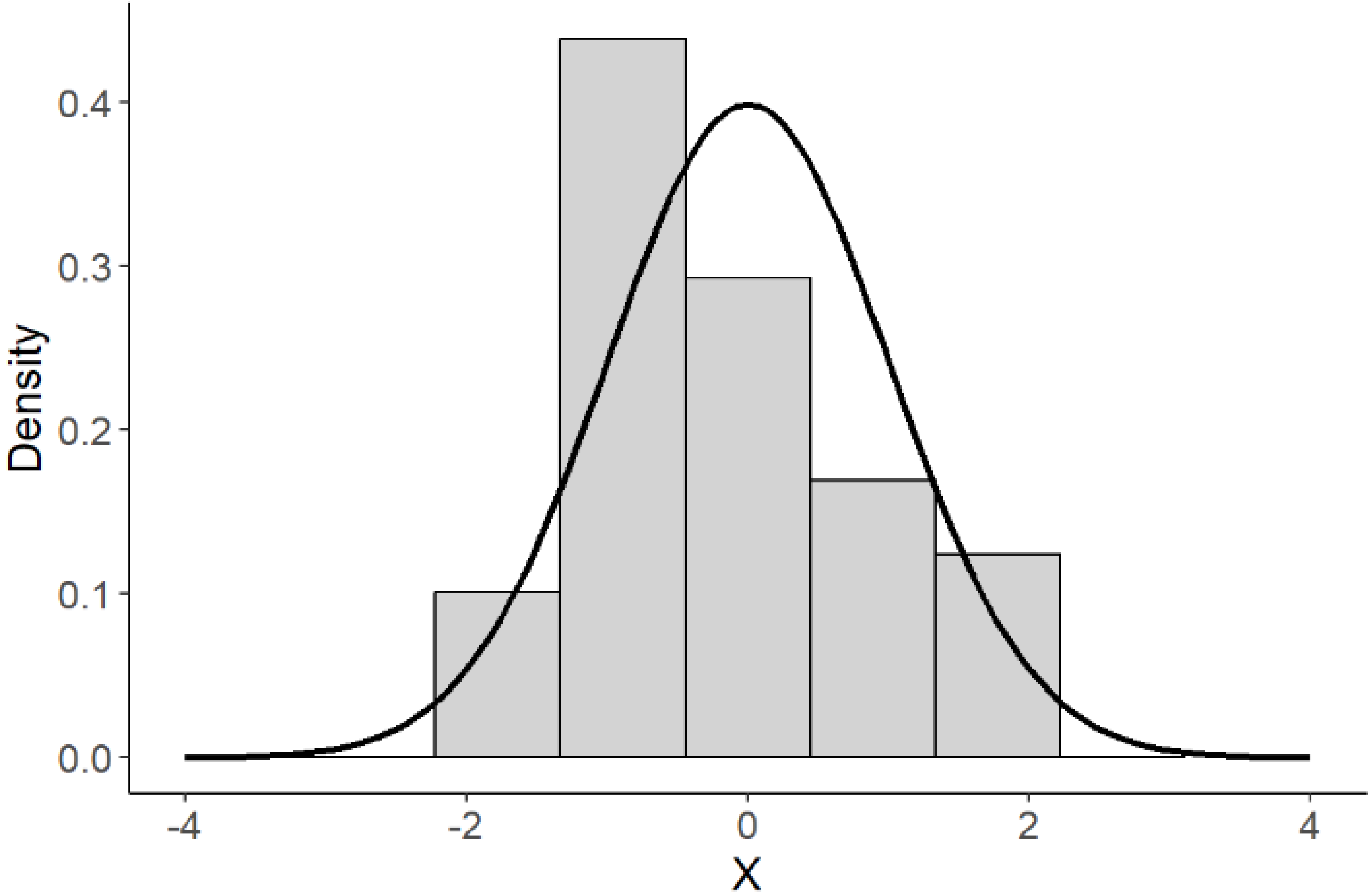
Lecture 7
The normal distribution,
Z-scores,
Transformations of Variables

Density Curves



The Normal Distribution

- A family of smooth, bell-shaped (symmetric) distributions that arise often in statistics

- Shape is determined by two parameters: the mean and the standard deviation

  The mean is located where the (relative) frequency is at its peak.

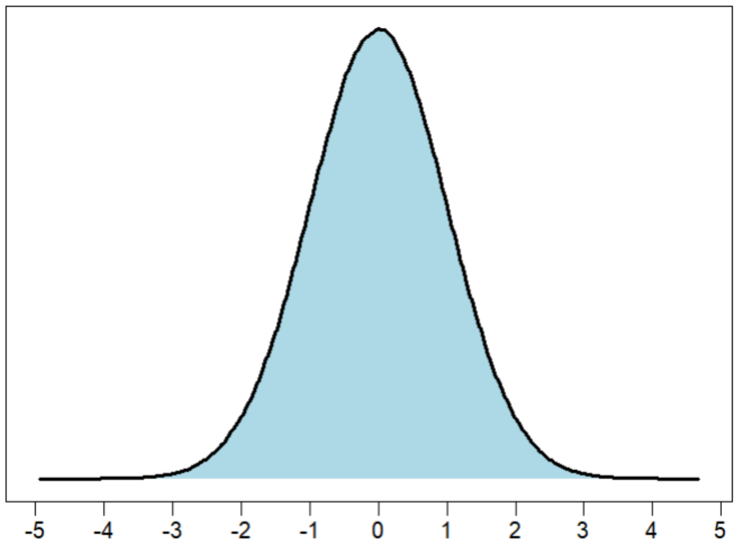
  The standard deviation is the distance from the mean to the value of the variable where the (relative) frequency is a little less than 3/4 of the way (actually about 68%) to its maximum.

- We denote the normal distribution for a population as

$$x \sim N(\mu, \sigma)$$

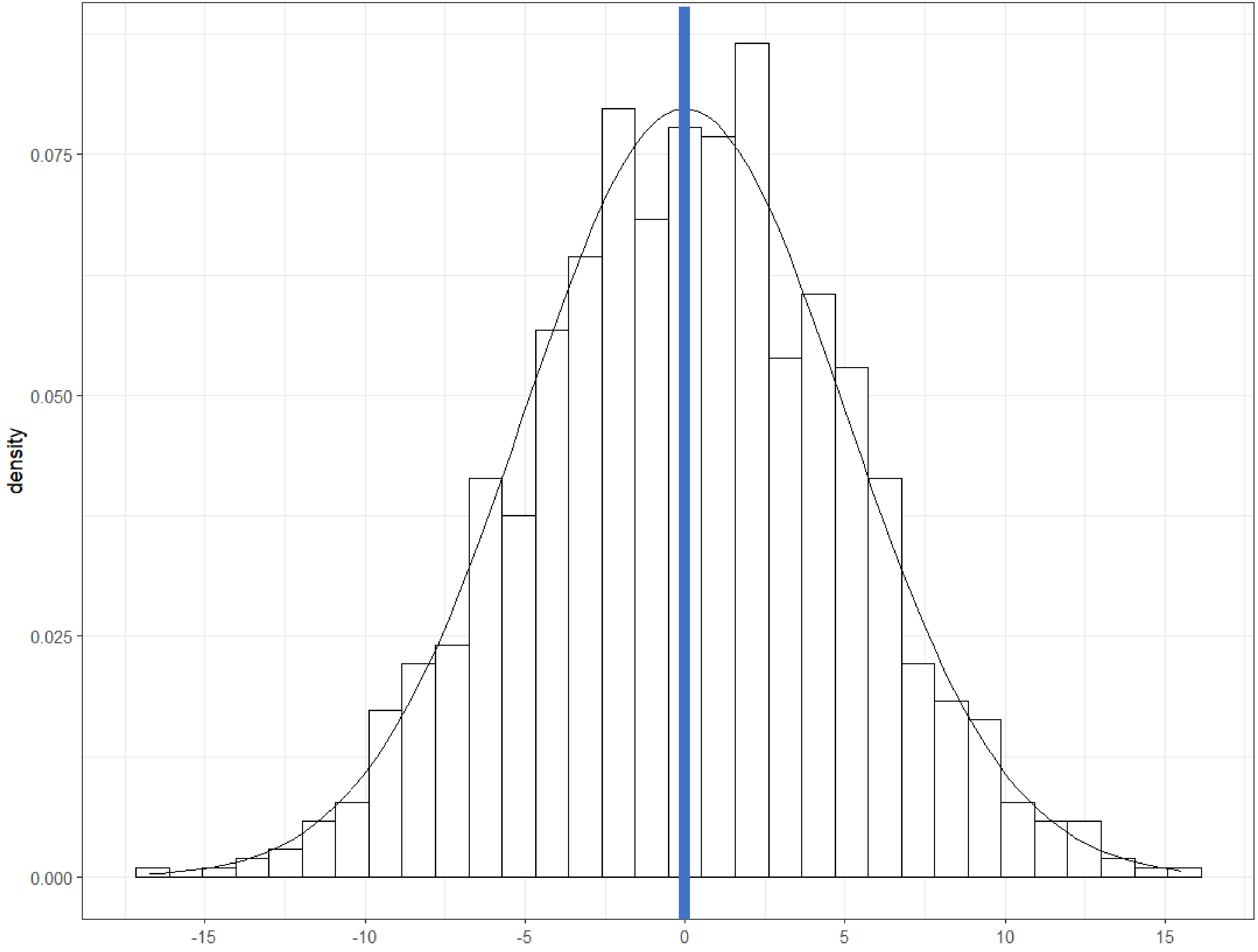
- And for a sample as

$$x \sim N(\bar{x}, s)$$



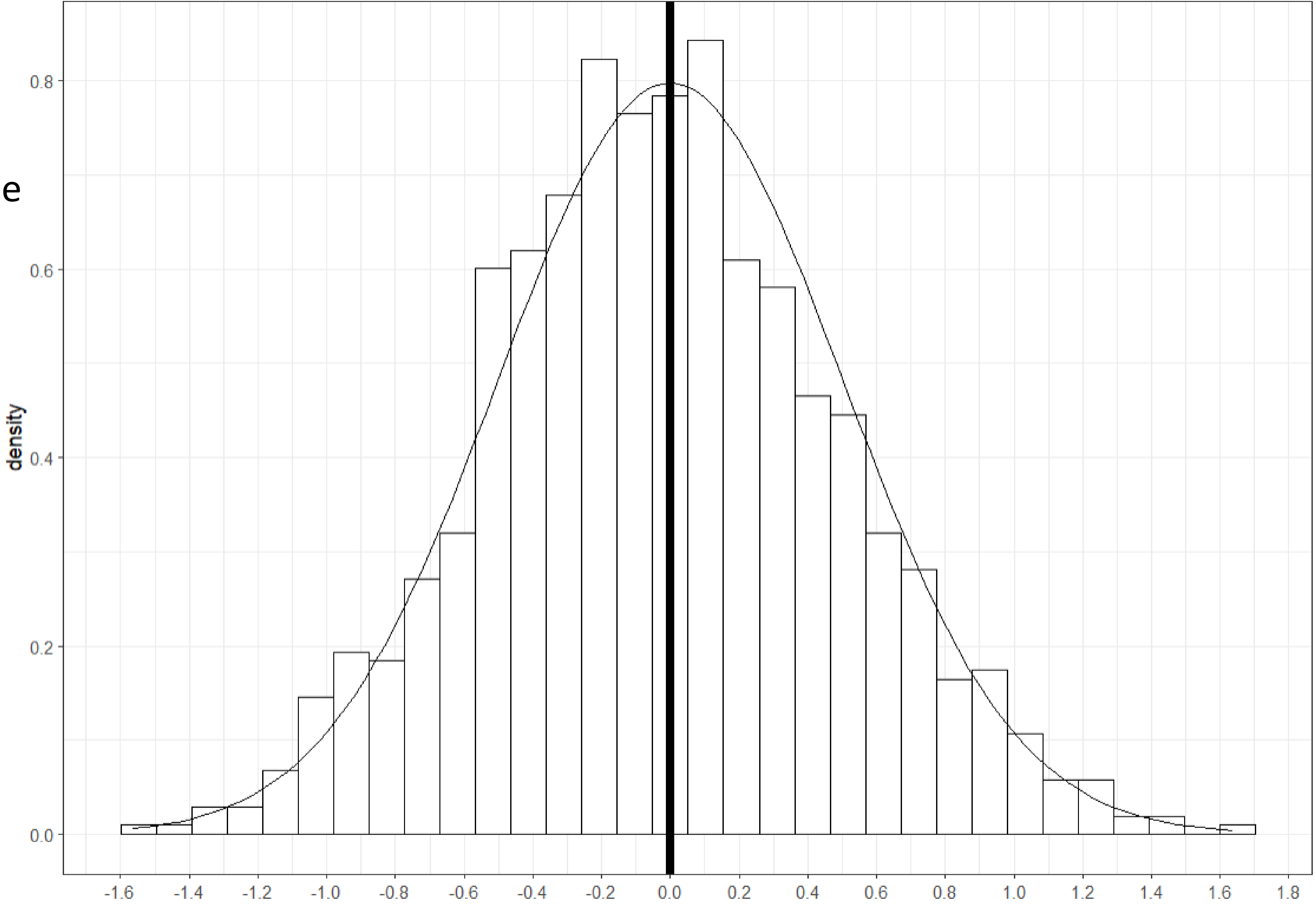
What is the approximate value of σ ?

$\sigma = 5$

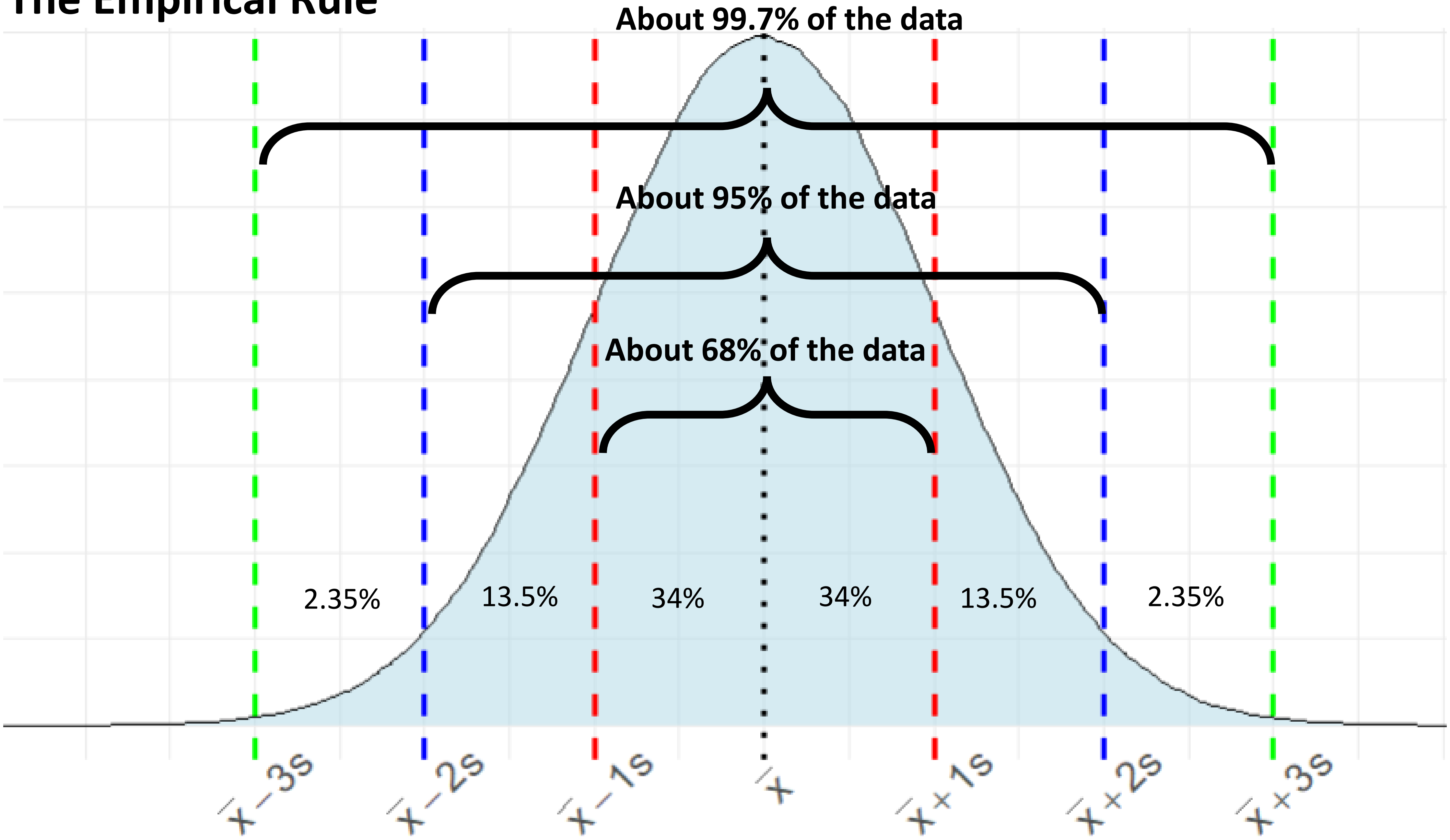


What is the approximate value of σ ?

$\sigma = 0.5$

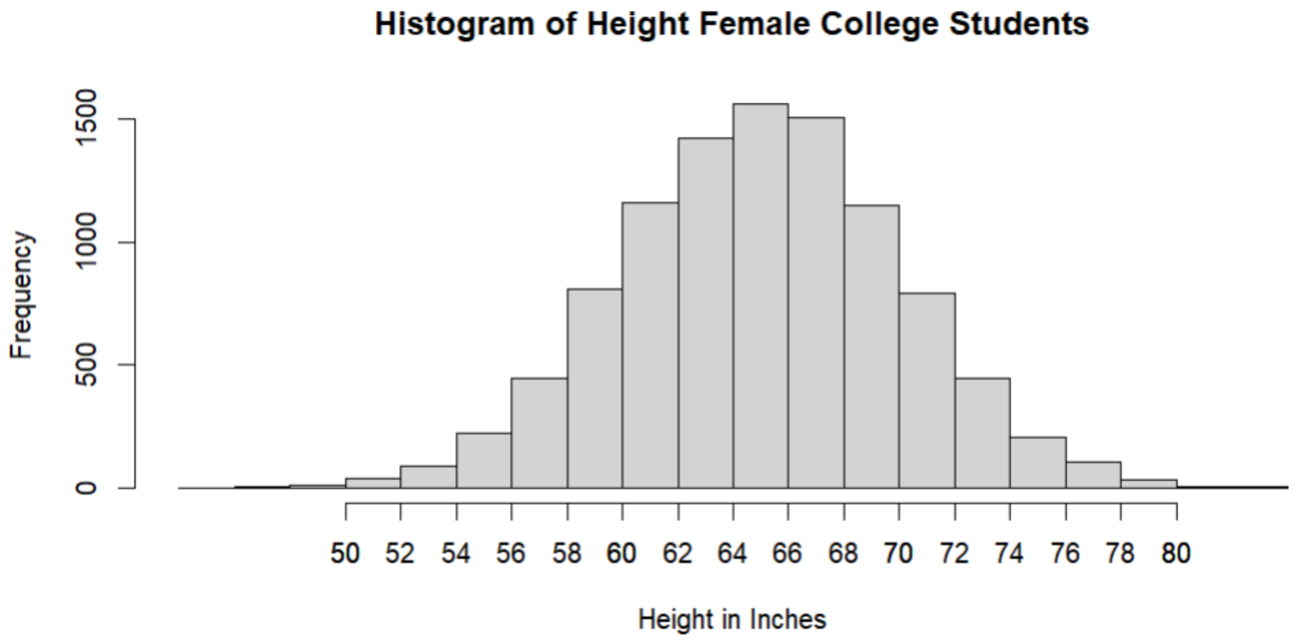


The Empirical Rule



Practice

- Suppose the distribution to the left represents the heights of a sample of female college students in the U.S. this distribution has mean and standard deviation
- $\bar{x} \approx 65$ inches
- $s \approx 5$ inches

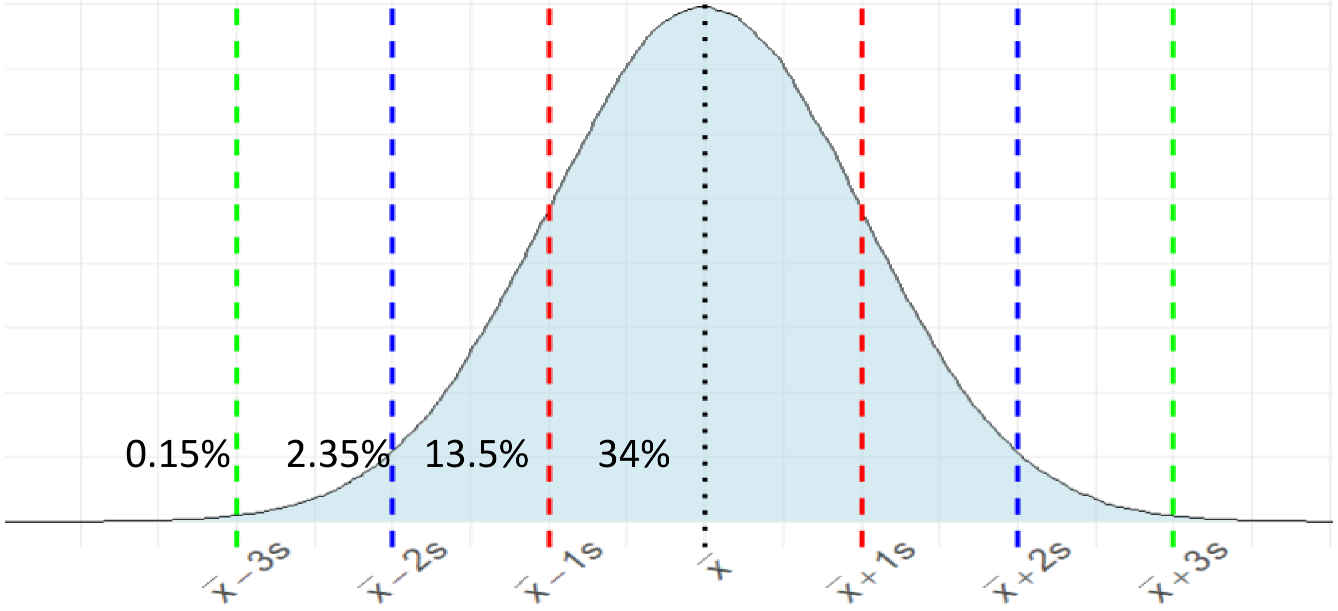


What percentage of students in the sample are shorter than mean height?

50%

What percentage of students in the sample are more than 2 standard deviations above the average height?

About 2.5%



Identifying Outliers: Normal Distributions

- The empirical rule: It is fairly unlikely to observe a value that is more than 2 standard deviations from the mean

- Therefore, when data are approximately Normally distributed, we can regard all values $\geq 2s$ distance from the mean as outliers

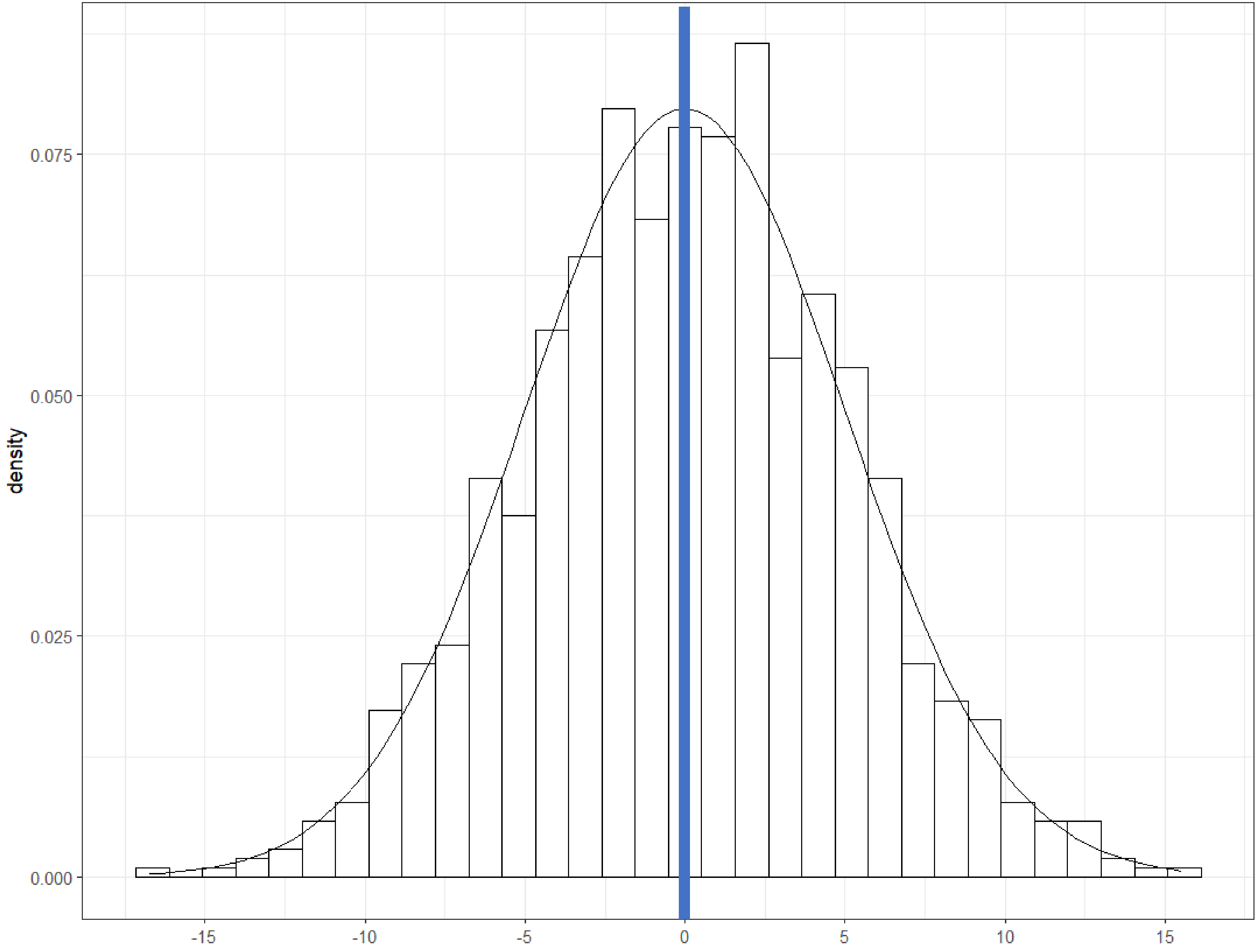
- $\boldsymbol{Z}$ **−score**: The number of standard deviations a value falls from mean

$$z_i = \frac{observation\ - mean}{standard\ deviation}$$

$$= \frac{x_i - \bar{x}}{s} \sim N(0,1) \text{ if } x_i \sim N(\mu, \sigma)$$

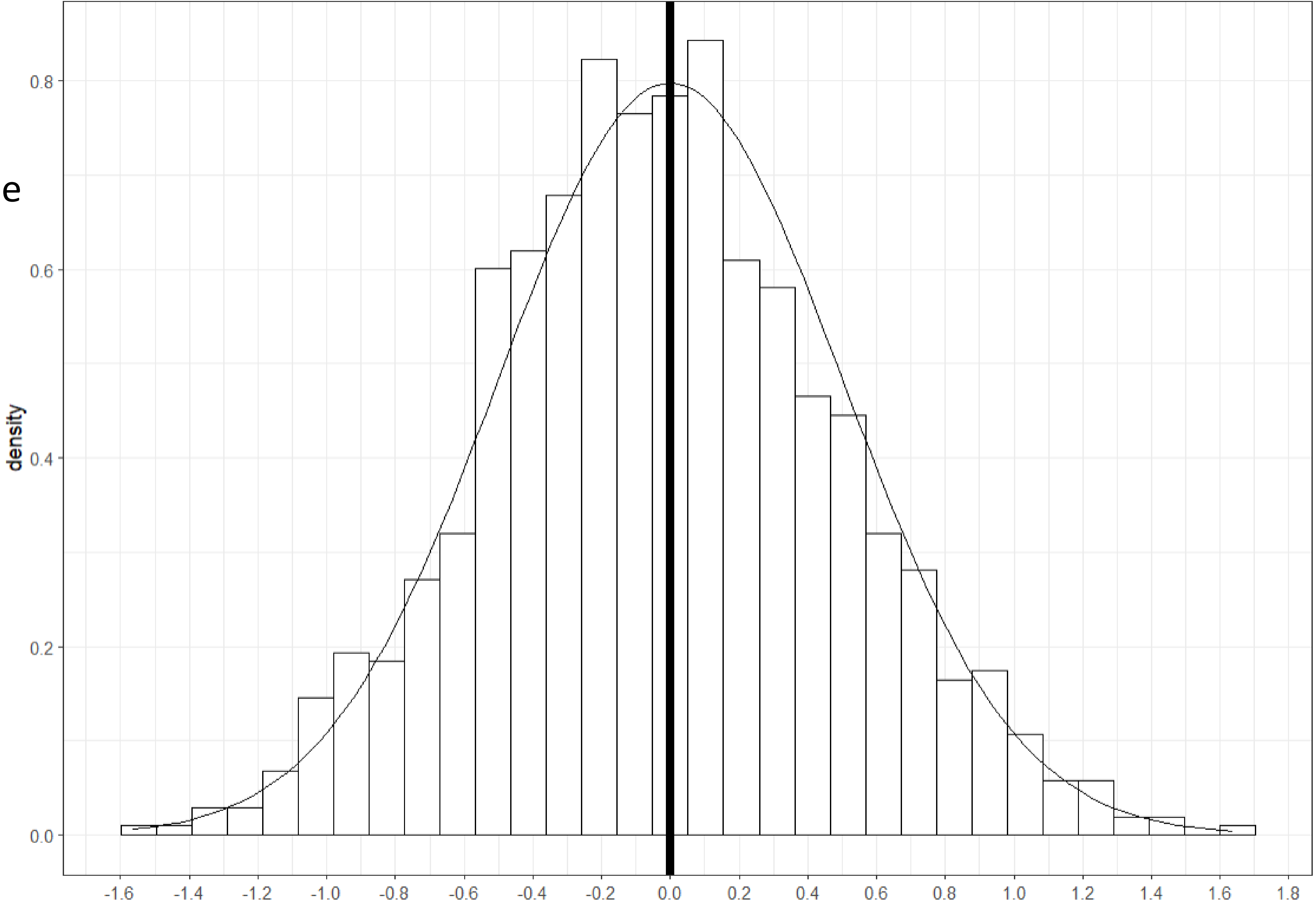
What is the approximate value of σ ?

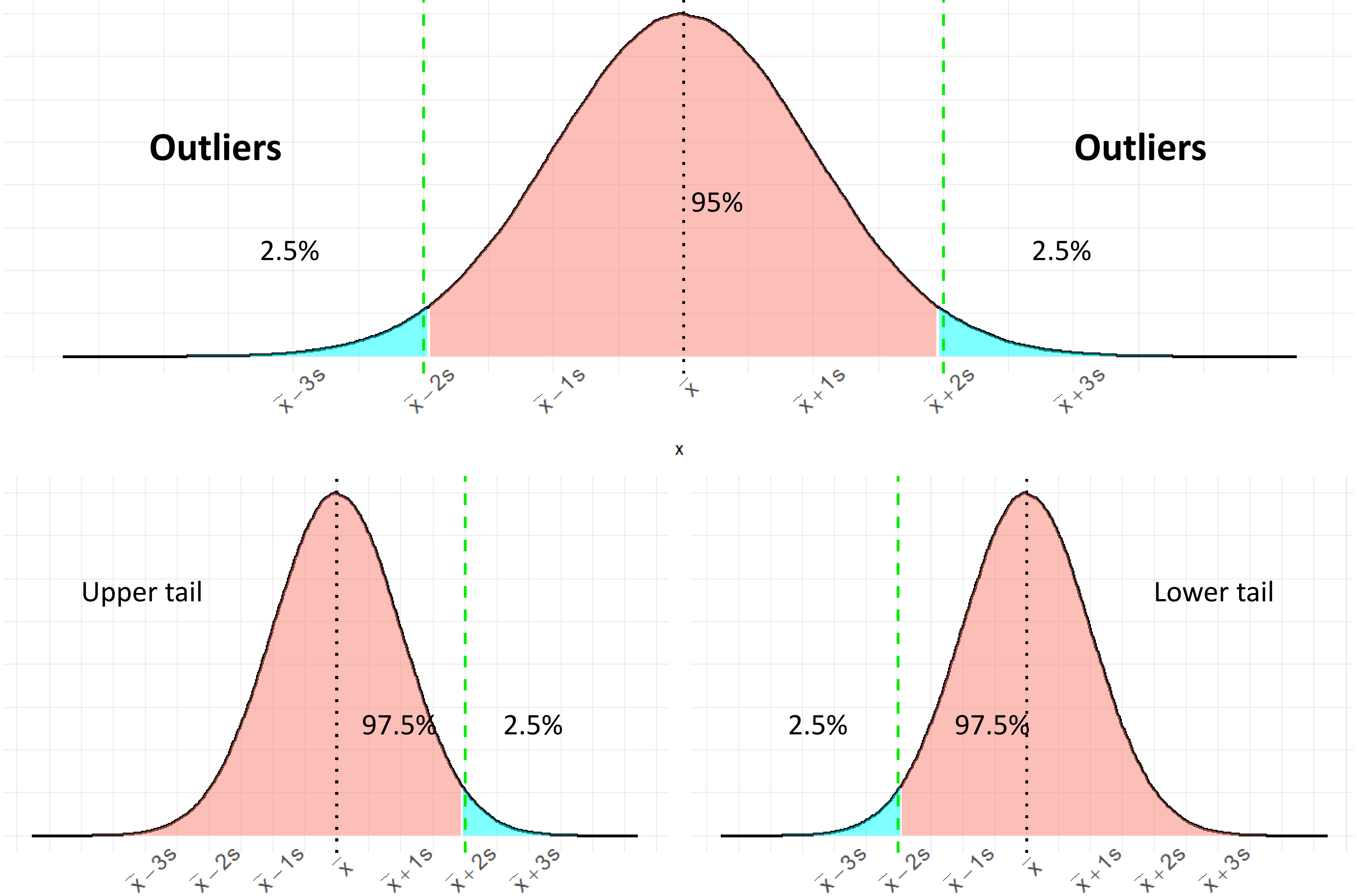
$\sigma = 5$



What is the approximate value of σ ?

$\sigma = 0.5$




Outliers
Outliers
95%
2.5%
2.5%
$\bar{x}-3s$
$\bar{x}-2s$
$\bar{x}-1s$
$\bar{x}$
$\bar{x}+1s$
$\bar{x}+2s$
$\bar{x}+3s$
x
Upper tail
97.5%
2.5%
$\bar{x}-3s$
$\bar{x}-2s$
$\bar{x}-1s$
$\bar{x}$
$\bar{x}+1s$
$\bar{x}+2s$
$\bar{x}+3s$
Lower tail
2.5%
97.5%
$\bar{x}-3s$
$\bar{x}-2s$
$\bar{x}-1s$
$\bar{x}$
$\bar{x}+1s$
$\bar{x}+2s$
$\bar{x}+3s$

Try it out: Female College Student Heights

Height	F(x)	RF(x)	CRF(x)
56	1	0.0038	0.0038
57	1	0.0038	0.0076
58	1	0.0038	0.0115
60	7	0.0267	0.0382
61	10	0.0382	0.0763
62	25	0.0954	0.1718
63	20	0.0763	0.2481
64	45	0.1718	0.4198
65	29	0.1107	0.5305
66	40	0.1527	0.6832
67	31	0.1183	0.8015
68	21	0.0802	0.8817
69	12	0.0458	0.9275
70	5	0.0191	0.9466
71	3	0.0115	0.9580
72	8	0.0305	0.9885
76	1	0.0038	0.9924
77	1	0.0038	0.9962
92	1	0.0038	1.0000

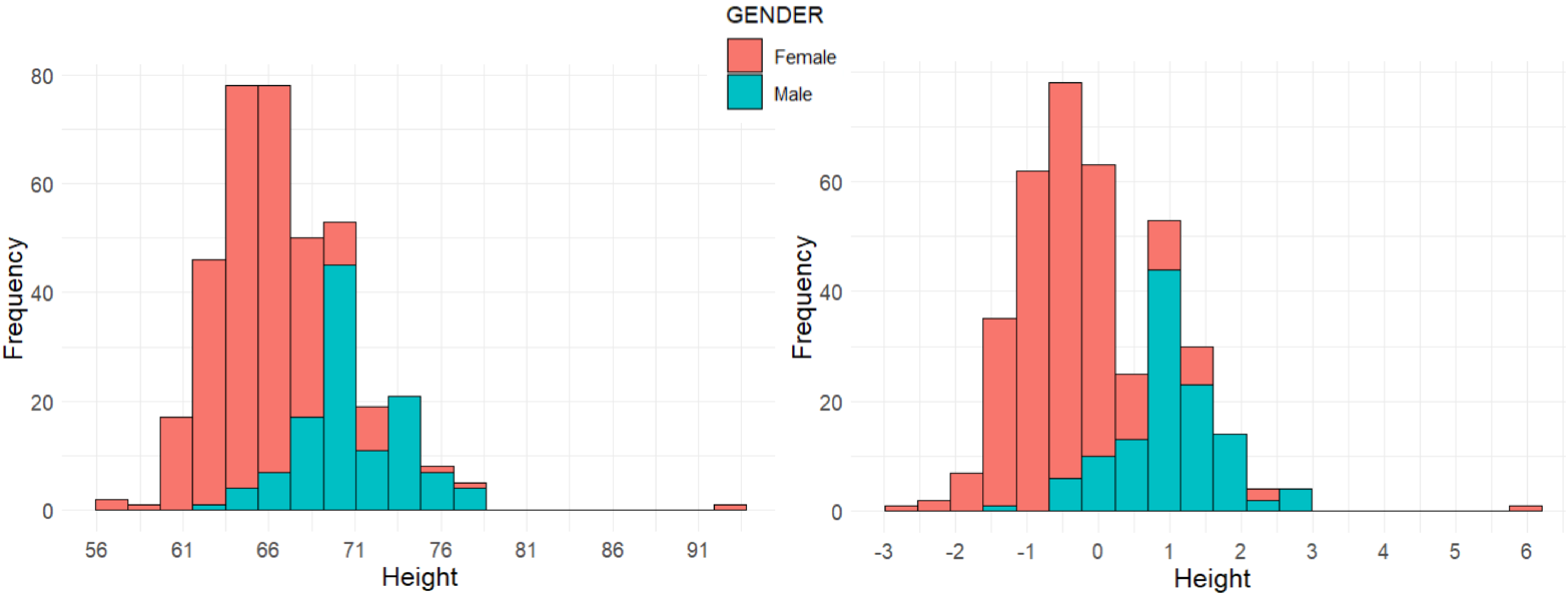
Compute the z-score for a female with a height of 70 inches

Compute the z-score for a female with a height of 92 inches

$$\bar{x} = 65.4$$

$$s = 3.38$$

College Student Heights



A Note About Transformations of Variables…

- We often need to change the units of measurement of a variable such as from Fahrenheit to Celsius, Feet to meters, dollars to euros etc.

- Linear transformations: adding, subtracting, multiplying, dividing
  - Linear transformations take the form $y = ax + b$ (scaling + shift)
  - a is a scaling constant, b is a shifting constant, x is the original variable and y the transformed variable
  - The z −score is a linear transformation
  - Linear transformations preserve the shape of variables distribution

- Nonlinear transformations: squaring, taking roots, logarithm, exponentiation, etc
  - **<u>Do not</u>** preserved the shape of the variables distribution

More properties of Linear Transformations

- For a linear transformation of x to y: $y = ax + b$
- $\bar{y} = a\bar{x} + b$
- $median(y) = a \cdot median(x) + b$
- $s_y = |a| \cdot s_x$ (the standard deviation is not affected by shift b)
- $IQR_y = |a| \cdot IQR_x$ (the IQR is not affected by shift b)

Identifying Outliers: Normal Distributions

- all values $\geq 2s$ distance from the mean are outliers

- $\boldsymbol{Z}$ $-$**score**: The number of standard deviations a value falls from mean

$$z_i = \frac{observation\ - mean}{standard\ deviation}$$

$$= \frac{x_i\ -\ \bar{x}}{s} \sim N(0,1)$$

Try it out: Female College Student Heights

Height	F(x)	RF(x)	CRF(x)
56	1	0.0038	0.0038
57	1	0.0038	0.0076
58	1	0.0038	0.0115
60	7	0.0267	0.0382
61	10	0.0382	0.0763
62	25	0.0954	0.1718
63	20	0.0763	0.2481
64	45	0.1718	0.4198
65	29	0.1107	0.5305
66	40	0.1527	0.6832
67	31	0.1183	0.8015
68	21	0.0802	0.8817
69	12	0.0458	0.9275
70	5	0.0191	0.9466
71	3	0.0115	0.9580
72	8	0.0305	0.9885
76	1	0.0038	0.9924
77	1	0.0038	0.9962
92	1	0.0038	1.0000

Compute the z-score for a female with a height of 70 inches

Compute the z-score for a female with a height of 92 inches

Assuming the distribution of the sample is approximately symmetric, about proportion students have a height between ?

How short does a female have to be before she would be considered an outlier relative to the data?

$$\bar{x} = 65.4$$

$$s = 3.4$$