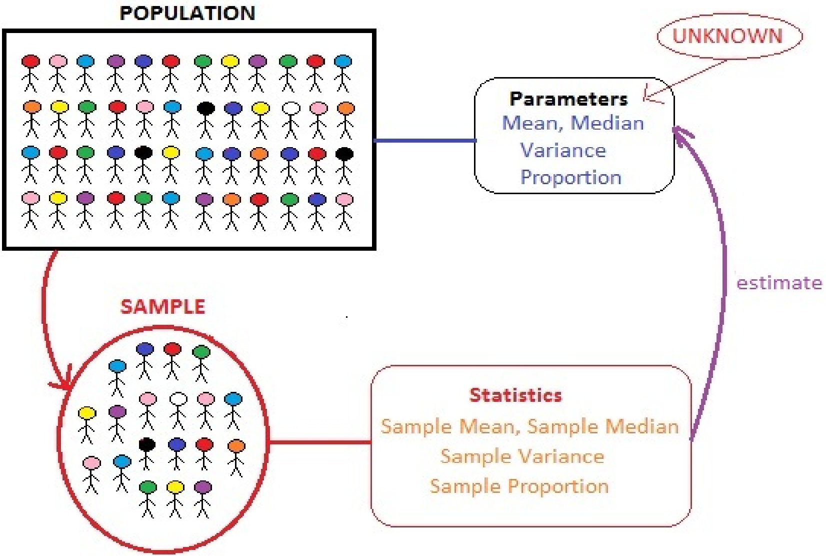
Descriptive Statistics

Statistical Inference

- Descriptive statistics: describe the data that we have collected
  - The Sample
- Statistical Inference: makes generalizations about something larger
  - The Population

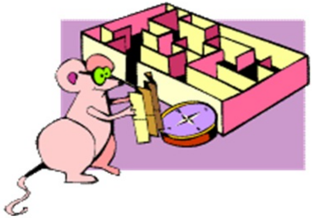


POPULATION
UNKNOWN
Parameters
Mean, Median
Variance
Proportion
estimate
SAMPLE
Statistics
Sample Mean, Sample Median
Sample Variance
Sample Proportion

Experiment

25 rats run through maze in an average of 34.5 seconds.

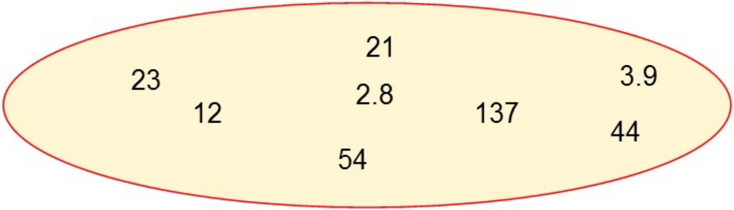
- Sample: those rats running that race
- Population: not only every rat but also every race they might run



Describing Samples

- Where is it? What is its center?
- What is the spread or variability? How much noise is in the data?
- What is the shape of the distribution? Is it symmetric?

Measuring these attributes



Center of the Distribution

Measures of the Center of the Distribution

- Mean: add up the data and divide by the number of observations.
- Median: An equal number of observations more and less than the median.

Mean

- Add up the data and divide by the number of observations

Examples

Data: 1, 2, 2, 3, 4

Mean = (1 + 2 + 2 + 3 + 4)/5 = 2.4

Data: 10, 12, 56, 78, 113, 1209

Mean = (10 + 12 + 56 + 78 + 113 + 1209)/6 = 246.3

Mean

- Data

$$\{x_1,\ x_2,\ x_3,\ \ldots, x_n\}$$

- Mean

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median

- The middle observation

  Data: 1, 2, **2**, 3, 4

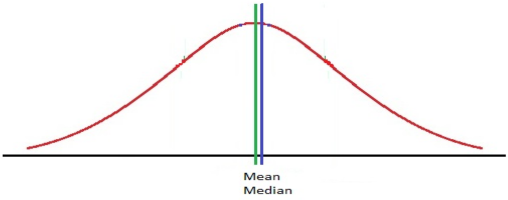
  Median = 2

  Data: 10, 12, **56, 78**, 113, 1209

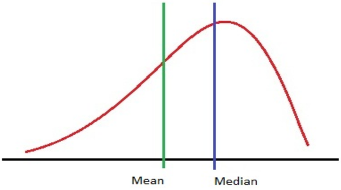
  Median = (56+78)/2 = 67

Mean versus Median

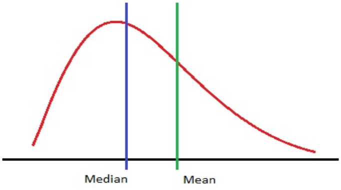
- The mean and the median are close for symmetric distributions.



- Mean moves in the direction of a skewed distribution



Skewed left



Skewed right

Outliers

- **Outlier** = a number that doesn't fit with the rest
- Data: 3, 6, 7, 10, **157**

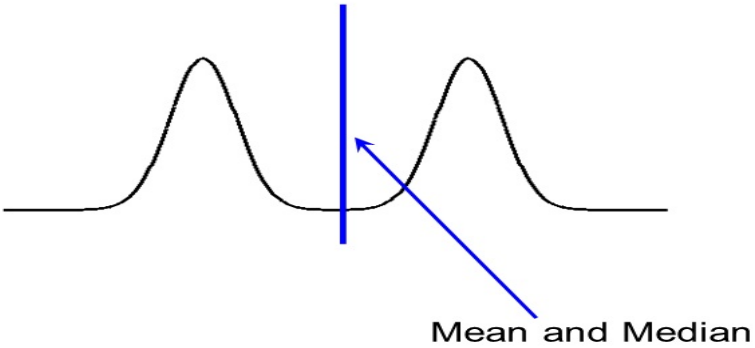
  Mean = $\frac{1}{5}(3+6+7+10+157) = 36.6$

  Median = 7

- Medians are resistant to Outliers.

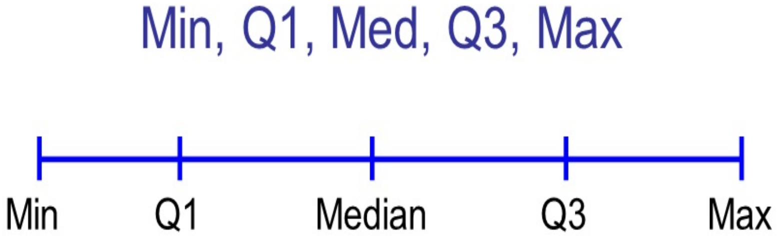
Modes

- **Mode**: peak in the distribution
- Bimodal = Two modes



5 number Summary

- Median
- Minimum, Maximum
- **Quartiles**: middle observation above the median and below the median



Finding Quartiles

Data: $7, 23, 75, 82, 34, 91, 10$

1. Sort:
   - 7, 10, 23, 34, 75, 82, 91
2. Find the median: 34
3. Below the median: 7, 10, 23
   - Lower Quartile $Q1 = 10$
4. Above the median: 75, 82, 91
   - Upper Quartile $Q3 = 82$

More Quartiles

Data: $7, 8, 22, 38, 48, 62$

1. Sort
2. Median = (22+38)/2 = 30
3. 7, 8, 22, 38, 48, 62
   - Lower Quartile: 7, 8, 22

     $$Q_1 = 8$$

   - Upper Quartile: 38, 48, 62

     $$Q_3 = 48$$

5 Number Summary

$$(min, Q1, med, Q3, max)$$

Example

$$\text{Data: } 1, 4, 5, 12, 34, 42, 56, 63, 71, 88$$

- Five Number Summary

$$(min, Q1, med, Q3, max) = (1, 5, 38, 63, 88)$$

Measuring the Spread

- How much variability is in the data?

1. Range = Maximum - Minimum
2. InterQuartile Range: $Q3 - Q1$
3. Standard Deviation: Average Square distance from the mean.

Sample Standard Deviation

- Formula

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Easier to calculate

$$s = \sqrt{\frac{1}{n-1}\left([\sum_{i=1}^{n} x_i^2] - n\bar{x}^2\right)}$$

5 Easy Steps

1. Calculate the mean $\bar{x}$.
2. Square it.
3. Calculate the sum of x^2.
4. Find the difference
$$(\text{sum of } x_i^2) - n\bar{x}^2$$
5. Divide by $n - 1$.
6. Take the square root.

Example: 7, 8, 3

1. Mean = $\bar{x}$ = 6
2. Mean squared = $\bar{x}^2$ = 36
3. (Sum of x^2) = $7^2 + 8^2 + 3^2 = 49 + 64 + 9 = 122$
4. (Sum of x^2) $- n\bar{x}^2$ = $122 - 3(36) = 122 - 108 = 14$
5. $\frac{1}{n-1}\, 14 = \frac{1}{3-1}\, 14 = 7$
6. $s = \sqrt{7} = 2.645$

IQR versus s

- IQR like the median does not depend on the largest (or smallest) observation (It is *outlier-resistant*).
- s depends on all data and can be sensitive to distant observations (outliers).

5 Number Summary

$$(\text{Minimum}, Q_1, \text{Median}, Q_3, \text{Maximum})$$

Example: 25, 78, 97, 133, 193, 212, 215, 274

- Median: (133+193)/2 = 163
- Lower part: 25, 78, 97, 133

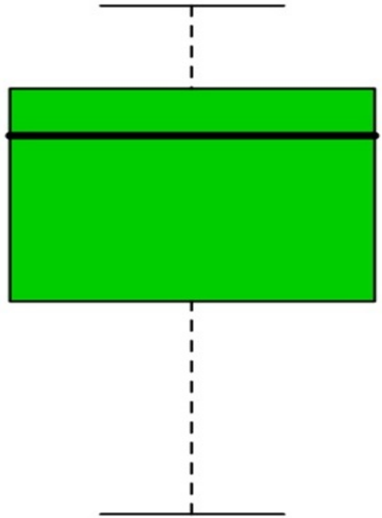
$$Q1 = (78 + 97)/2 = 87.5$$

- Lower part: 193, 212, 215, 274

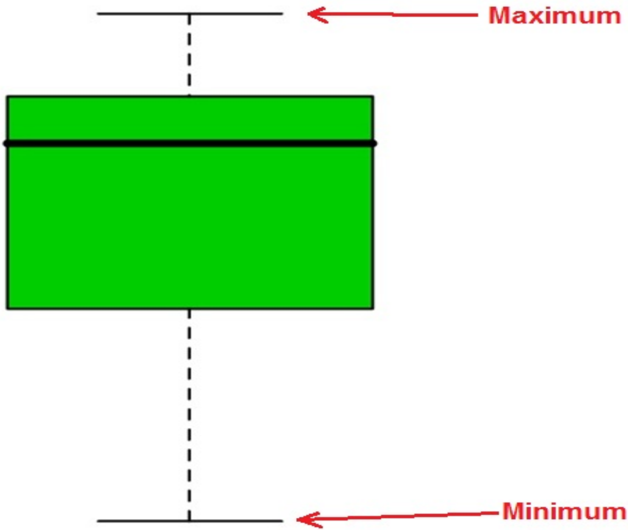
$$Q3 = (212 + 215)/2 = 213.5$$

$$(25, 87.5, 163, 213.5, 274)$$

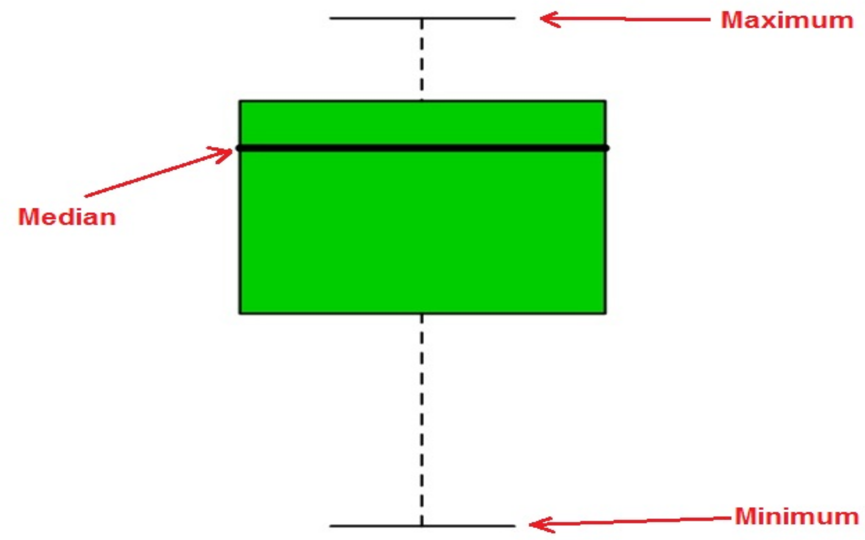
Box Pot



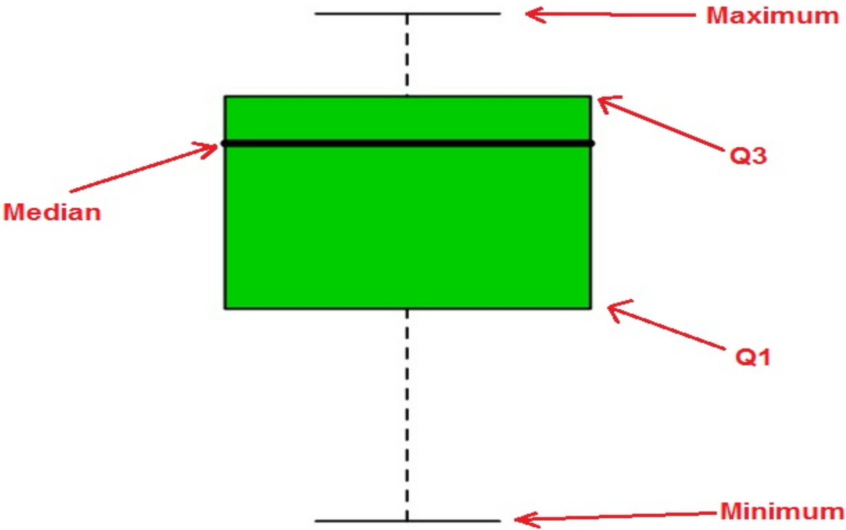
Box Pot



Box Pot



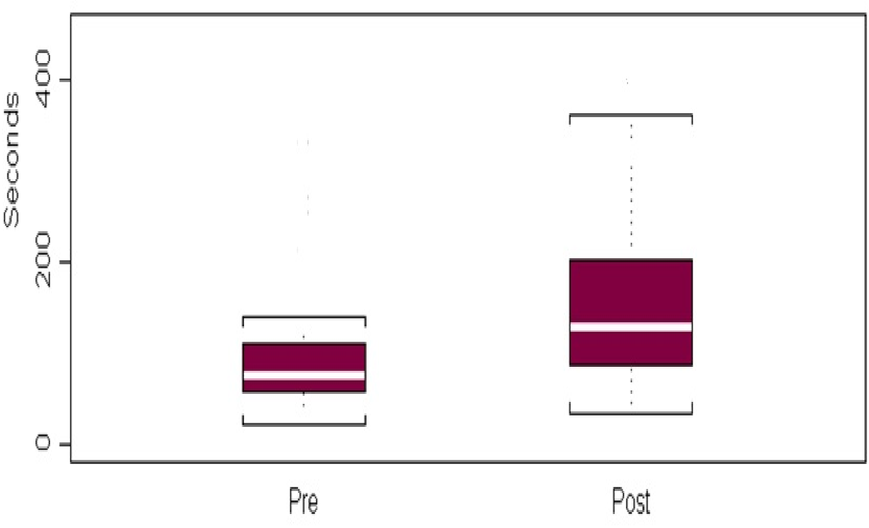
Box Pot



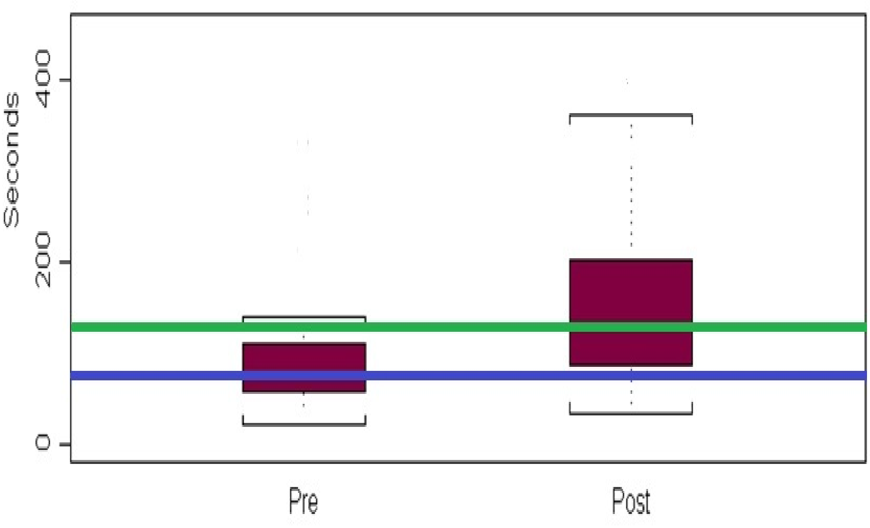
Groups' Comparison

- Side-by-side boxplots to compare two or more sets of data.
  - Do they have the same center? Shape? Spread?
  - Is the difference between the medians much bigger than the variability in the data?

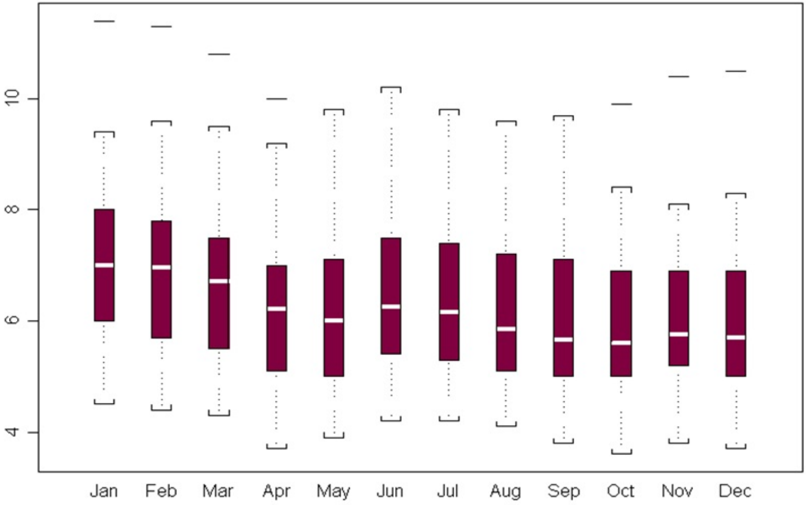
Pulmonary test scores pre/post treatment



Pulmonary test scores pre/post treatment

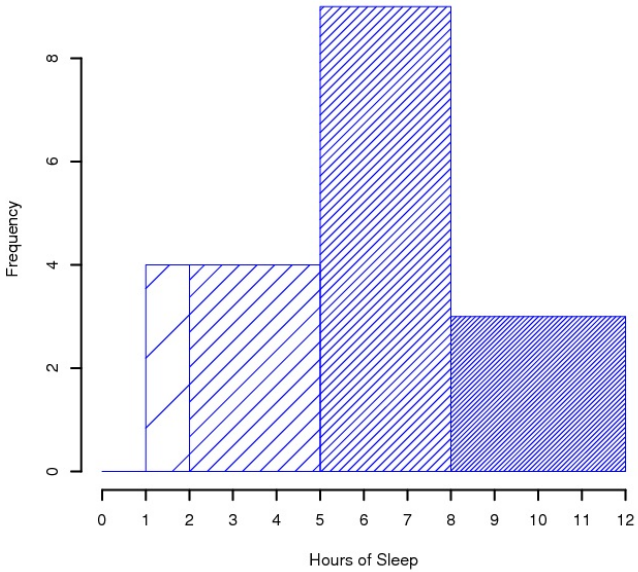


Seasonal behavior of unemployment rates



Data from 1980–2001

The Histogram



The Histogram

Example: *Hours of sleep*

$$\Big\{12, 8.5, 7.2, 7.3, 7.7, 6, 6.5, 4.5, 3, 1.2,$$
$$1.3, 2, 2, 3.8, 6.6, 8.5, 5.9, 4.6, 5.6, 6.7\Big\}$$

Variable: `hours of sleep`,
Values = $[0, 24]$.

1. How many blocks are we going to have?
2. How are we going to determine the length of each block?

Example: *Hours of sleep*

1. Sort the data:

$$\{1.2, 1.3, 2, 2, 3, 3.8, 4.5, 4.6, 5.6, 5.9,$$
$$6, 6.5, 6.6, 6.7, 7.2, 7.3, 7.7, 8.5, 8.5, 12\}$$

2. Choose the desired *class intervals*:

   1-2 hours, 2-5 hours, 5-8 hours, 8-12 hours

   - 4 class intervals
   - 4 unevenly spaced blocks

Example: *Hours of sleep*

How to draw the block?

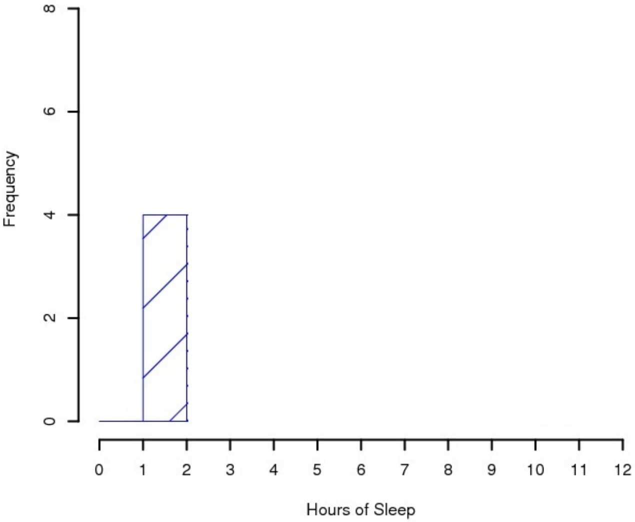
- Count the number of datapoints that falls into each class:

Hours of Sleep (X)	*Counts*	*Proportions*
$1 < X \leq 2$	4	4/20=0.2
$2 < X \leq 5$	4	4/20=0.2
$5 < X \leq 8$	9	9/20=0.45
$8 < X \leq 12$	3	3/20=0.15

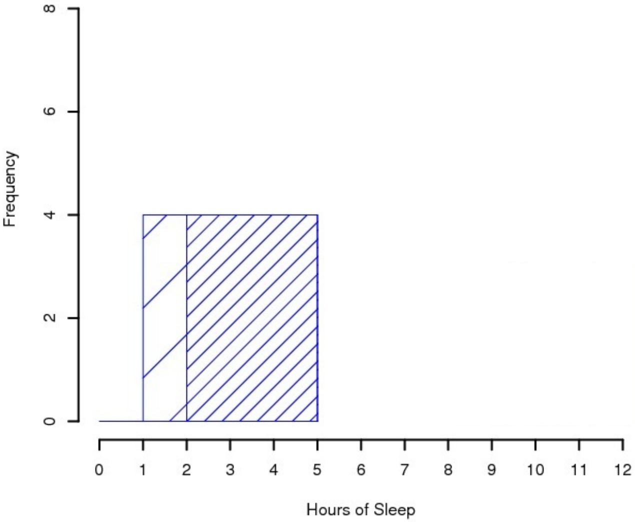
The intervals are not necessary to have the same length.

Example: *Hours of sleep*

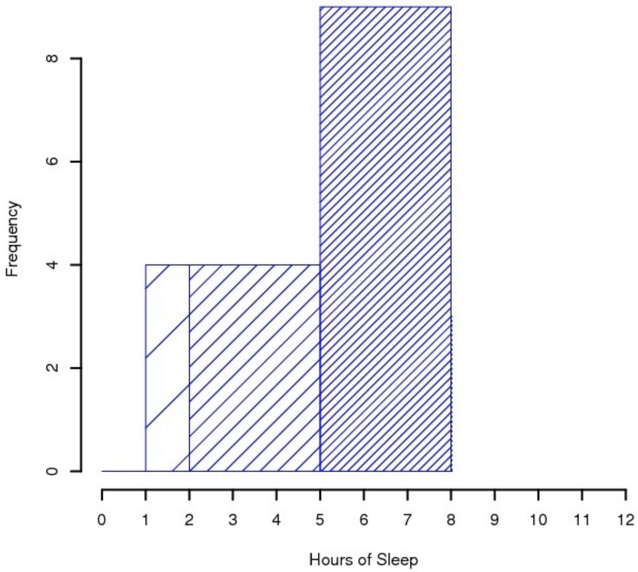
Example: *Hours of sleep*



Example: *Hours of sleep*



Example: *Hours of sleep*



Example: *Hours of sleep*