Why are data collected?

What kind of data are collected?

What kind of data are collected?

- Medical & pharmaceutical sciences...

What kind of data are collected?

- Medical & pharmaceutical sciences...
- Financial companies & institutions...

What kind of data are collected?

- Medical & pharmaceutical sciences...
- Financial companies & institutions...
- Engineering & computer science...

What kind of data are collected?

- ❑ Medical & pharmaceutical sciences...
- ❑ Financial companies & institutions...
- ❑ Engineering & computer science...
- ❑ Social & behavioral sciences...

What kind of data are collected?

- Medical & pharmaceutical sciences...
- Financial companies & institutions...
- Engineering & computer science...
- Social & behavioral sciences...
- Education & research...

What kind of data are collected?

- Medical & pharmaceutical sciences...
- Financial companies & institutions...
- Engineering & computer science...
- Social & behavioral sciences...
- Education & research...
- Sport, entertainment & fun...

What kind of data are collected?

- Medical & pharmaceutical sciences...
- Financial companies & institutions...
- Engineering & computer science...
- Social & behavioral sciences...
- Education & research...
- Sport, entertainment & fun...
- etc.

How to do it in a proper way?

Syllabus

□ Introduction & Descriptive Statistics	*chapter 6*
□ Introduction to Probability	*chapter 1*
□ Random Variables	*chapter 2*
□ Discrete and Continuous Probability Distributions	*chapters 3,4*
□ The Normal Probability Distribution	*chapter 5*
□ Sampling Distributions, Random Sample	*chapter 7*
□ Inferences on a Population Mean	*chapter 8*
□ Comparing Two Population Means	*chapter 9*
□ Simple Linear Regression and Correlation	*chapter 12*
□ Inferences on a Population Proportion	*chapter 10*
□ The Analysis of Variance	*chapter 11*

How to do it in a proper way?

How to do it in a proper way?

Probability $\overset{}{\underset{\text{vs.}}{\longleftrightarrow}}$ Statistics

- ❑ Random mechanism → produces random outcomes;

How to do it in a proper way?

Probability $\overset{}{\underset{\text{vs.}}{\longleftrightarrow}}$ **Statistics**

- ❑ Random mechanism → produces random outcomes;
- ❑ A set of random outcomes (results) → drawing conclusions;

How to do it in a proper way?

Probability $\xleftrightarrow[\text{vs.}]{}$ Statistics

- ❑ Random mechanism → produces random outcomes;
- ❑ A set of random outcomes (results) → drawing conclusions;

- ❑ **Probability** - theoretical background for a random mechanism; *(describes how the mechanism works - but the principle is unknown)*

How to do it in a proper way?

Probability $\underset{\text{vs.}}{\longleftrightarrow}$ Statistics

- □ Random mechanism $\rightarrow$ produces random outcomes;
- □ A set of random outcomes (results) $\rightarrow$ drawing conclusions;

- □ **Probability** - theoretical background for a random mechanism;
  (describes how the mechanism works - but the principle is unknown)
- □ **Statistics** - uses random outcomes to draw conclusions;
  (using statistics we can very precisely describe the mechanism behind)

Statistics in Real Life...

- ❑ In probability, we could have a precise description of the random mechanism behind which would be accurate, but it is unknown!
- ❑ In statistics, we have data (measurements) that are quite often not precise, but the final conclusions drawn from such data are very accurate!

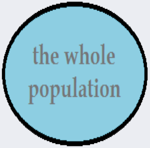
Statistics in Real Life...

- In probability, we could have a precise description of the random mechanism behind which would be accurate, but it is unknown!

- In statistics, we have data (measurements) that are quite often not precise, but the final conclusions drawn from such data are very accurate!

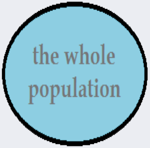
- "**I only believe in statistics that I doctored myself.**"
  Winston S. Churchill (1874 – 1965)

Population vs. Sample



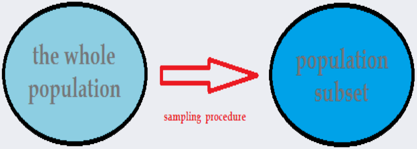
Population vs. Sample

Probability Theory
Random Distribution



Population vs. Sample

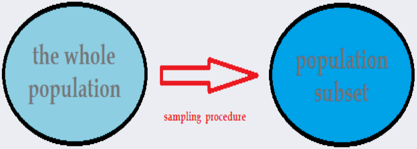
Probability Theory
Random Distribution



Population vs. Sample

Probability Theory
Random Distribution

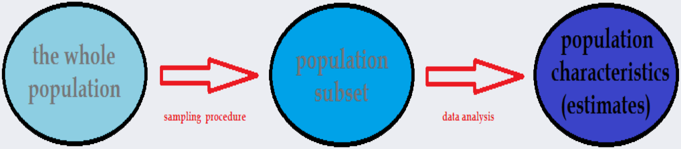
Finite Sample
Empirical Distribution



Population vs. Sample

Probability Theory
Random Distribution

Finite Sample
Empirical Distribution

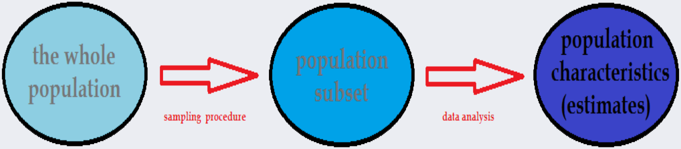


Population vs. Sample

Probability Theory
Random Distribution

Finite Sample
Empirical Distribution

Sample Statistics

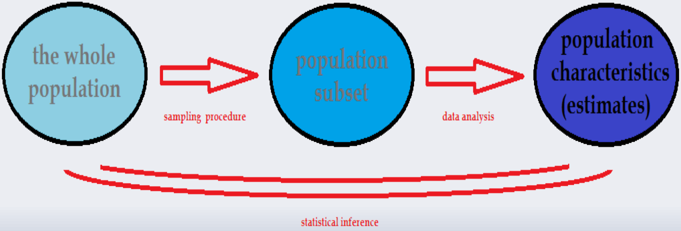


Population vs. Sample

Probability Theory
Random Distribution

Finite Sample
Empirical Distribution

Sample Statistics
Statistical Inference



Different Data types

- **Categorical variables**

Different Data types

- ❑ **Categorical variables**

- ❑ **Numerical variables**

Different Data types

- **Categorical variables**
  - Nominal categories
    - categories with no ordering;
      What kind of transportation do you use to get to work?
      What kind of material do you mostly prefer?
- **Numerical variables**

Different Data types

- **Categorical variables**
  - Nominal categories
    - categories with no ordering;
      What kind of transportation do you use to get to work?
      What kind of material do you mostly prefer?
  - Ordinal categories
    - categories with possible ordering;
      What was your most frequent grade last semester?
      What energy category does the machine belong to?
- **Numerical variables**

Different Data types

- ❑ **Categorical variables**
  - ❑ Nominal categories
    - ❑ categories with no ordering;
      What kind of transportation do you use to get to work?
      What kind of material do you mostly prefer?
  - ❑ Ordinal categories
    - ❑ categories with possible ordering;
      What was your most frequent grade last semester?
      What energy category does the machine belong to?
- ❑ **Numerical variables**
  - ❑ Integer values - counts
    - ❑ an apriori assumption of equidistant differences;
      How many people visit this museum a day?
      How many call a day is addressed to 911?

Different Data types

- **Categorical variables**
  - Nominal categories
    - categories with no ordering;
      What kind of transportation do you use to get to work?
      What kind of material do you mostly prefer?
  - Ordinal categories
    - categories with possible ordering;
      What was your most frequent grade last semester?
      What energy category does the machine belong to?
- **Numerical variables**
  - Integer values - counts
    - an apriori assumption of equidistant differences;
      How many people visit this museum a day?
      How many call a day is addressed to 911?
  - Real values
    - any real value is possible...
      length, weight, height, temperature, etc.

Data exploration

- ❑ What is the nature of data?
- ❑ What are the limitations of the experiment behind?

- ❑ What is the main question of interest?
- ❑ Is it possible to use data available to answer it?

- ❑ Is it important to visualize the data?
- ❑ How can one get an insight in the data?

Binary Data

- the simplest data type...
- different coding possible...
  - logical: TRUE | FALSE; 1|0; YES|NO;
  - categorical: A|B; 1|2; HOME|ABROAD;
- only a few options on how to represent such data;

Binary Data

- the simplest data type...
- different coding possible...
  - logical: TRUE | FALSE; 1|0; YES|NO;
  - categorical: A|B; 1|2; HOME|ABROAD;
- only a few options on how to represent such data;

Summary (table):

TRUE	41
FALSE	59

Binary Data

- the simplest data type...
- different coding possible...
    - logical: TRUE | FALSE; 1|0; YES|NO;
    - categorical: A|B; 1|2; HOME|ABROAD;
- only a few options on how to represent such data;

Summary (table):

TRUE	41
FALSE	59

Frequency summary:

TRUE: 41.00 %

Binary Data

- the simplest data type...
- different coding possible...
  - logical: TRUE | FALSE; 1|0; YES|NO;
  - categorical: A|B; 1|2; HOME|ABROAD;
- only a few options on how to represent such data;

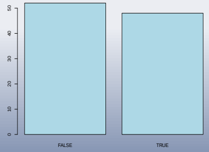
Summary (table):

TRUE	41
FALSE	59

Frequency summary:

TRUE: 41.00 %

Graphical view:



Categorical Data

- **Nominal categories...**
  - **Machine Breakdowns:** what is a machine breakdown cause?

Categorical Data

- **Nominal categories...**
  - **Machine Breakdowns:** what is a machine breakdown cause? *elektrical* | *mechanical* | *misuse*

Categorical Data

- **Nominal categories...**
  - **Machine Breakdowns:** what is a machine breakdown cause?
    elektrical | *mechanical* | *misuse*
- Data: a sequence of words 'electrical', 'mechanical' or 'misuse';

Categorical Data

- **Nominal categories...**
  - **Machine Breakdowns:** what is a machine breakdown cause?
    elektrical | *mechanical* | *misuse*
- Data: a sequence of words 'electrical', 'mechanical' or 'misuse';

Summary (table):

Electrical	9
Mechanical	24
Misuse	13

Categorical Data

- **Nominal categories...**
  - **Machine Breakdowns:** what is a machine breakdown cause? *elektrical* | *mechanical* | *misuse*
- Data: a sequence of words 'electrical', 'mechanical' or 'misuse';

Summary (table):

Electrical	9
Mechanical	24
Misuse	13

Frequency summary:

Electrical	19.57 %
Mechanical	52.17 %
Misuse	28.26 %

Categorical Data

- ❑ **Nominal categories...**
    - ❑ **Machine Breakdowns:** what is a machine breakdown cause?
      elektrical | *mechanical* | *misuse*
- ❑ Data: a sequence of words 'electrical', 'mechanical' or 'misuse';

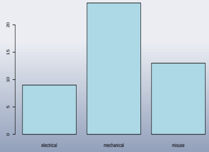
Summary (table):

Electrical	9
Mechanical	24
Misuse	13

Frequency summary:

Electrical	19.57 %
Mechanical	52.17 %
Misuse	28.26 %

Graphical view:



Categorical Data

- **Nominal categories...**
  - **Machine Breakdowns:** what is a machine breakdown cause?
    elektrical | *mechanical* | *misuse*
- Data: a sequence of words 'electrical', 'mechanical' or 'misuse';

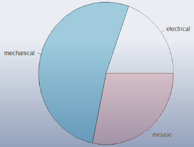
Summary (table):

Electrical	9
Mechanical	24
Misuse	13

Frequency summary:

Electrical	19.57 %
Mechanical	52.17 %
Misuse	28.26 %

Graphical view:



Categorical Data

- **Ordinal categories...**
  - **Number of Breakdowns:** How many times did a machine break down?

Categorical Data

- **Ordinal categories...**
  - **Number of Breakdowns:** How many times did a machine break down?
    0 times | *once* | *two times and more*

Categorical Data

- **Ordinal categories...**
  - **Number of Breakdowns:** How many times did a machine break down? *0 times | once | two times and more*
- Data: a sequence of values 0,1 and 2;

Categorical Data

- **Ordinal categories...**
  - **Number of Breakdowns:** How many times did a machine break down?
    0 times | once | two times and more
- Data: a sequence of values 0,1 and 2;

Summary (table):

Zero	54
Once	40
≥ 2	6

Categorical Data

- **Ordinal categories...**
  - **Number of Breakdowns:** How many times did a machine break down?
    0 times | *once* | *two times and more*
- Data: a sequence of values 0,1 and 2;

Summary (table):

Zero	54
Once	40
≥ 2	6

Frequency summary:

Zero	54 %
Once	40 %
≥ 2	6 %

Categorical Data

- **Ordinal categories...**
  - **Number of Breakdowns:** How many times did a machine break down? *0 times | once | two times and more*
- Data: a sequence of values 0,1 and 2;

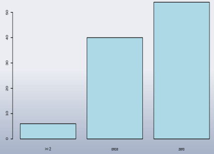
Summary (table):

Zero	54
Once	40
≥ 2	6

Frequency summary:

Zero	54 %
Once	40 %
≥ 2	6 %

Graphical view:



Categorical Data

- **Ordinal categories...**
  - **Number of Breakdowns:** How many times did a machine break down? *0 times | once | two times and more*
- Data: a sequence of values 0,1 and 2;

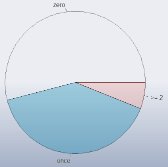
Summary (table):

Zero	54
Once	40
≥ 2	6

Frequency summary:

Zero	54 %
Once	40 %
≥ 2	6 %

Graphical view:



Categorical Data

- **Ordinal categories...**
  - **Number of Breakdowns:** How many times did a machine break down?
    0 times | once | two times and more
- Data: a sequence of values 0,1 and 2;

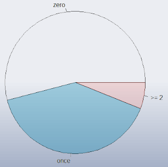
Summary (table):

Zero	54
Once	40
≥ 2	6

Frequency summary:

Zero	54 %
Once	40 %
≥ 2	6 %

Graphical view:



- What is the main difference between nominal and ordinal categories?

Numerical Data

- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?

Numerical Data

- **Integer values - counts...**
  - How many hours can a machine work until it breaks down? *0, 1, 2, 3, ... etc.*

Numerical Data

- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.

Numerical Data

- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?

Numerical Data

- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?
    1, 2, 3, etc. ≤ 5

Numerical Data

- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?
    1, 2, 3, etc. ≤ 10

Numerical Data

- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?
    1, 2, 3, etc. ≤ 30

Numerical Data

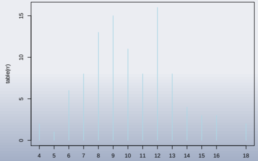
- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?
    1, 2, 3, etc. ≤ 100

Numerical Data

- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?
    1, 2, 3, etc. ≤ 100
- Data: mostly a sequence of integer values...

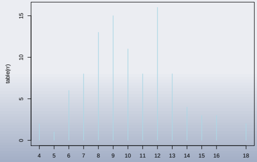
Numerical Data

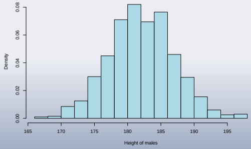
- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?
    1, 2, 3, etc. ≤ 100
- Data: mostly a sequence of integer values...



Numerical Data

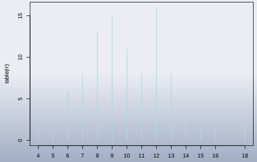
- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?
    1, 2, 3, etc. ≤ 100
- Data: mostly a sequence of integer values...

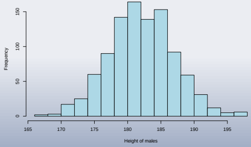




Numerical Data

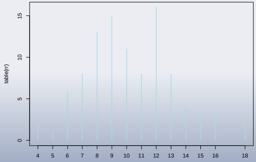
- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?
    1, 2, 3, etc. ≤ 100
- Data: mostly a sequence of integer values...

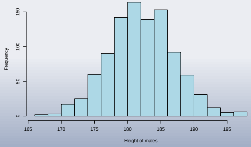


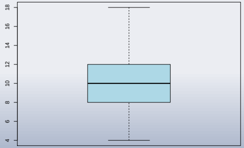


Numerical Data

- **Integer values - counts...**
  - How many hours can a machine work until it breaks down?
    0, 1, 2, 3, ... etc.
  - How many people a day will use the lift?
    0, 1, 2, 3, ... etc.
  - How many passengers is in a driving car?
    1, 2, 3, etc. ≤ 100
- Data: mostly a sequence of integer values...







Numerical Data

- **Real values...**
  - We could improve a precision on the previous example...

Numerical Data

- **Real values...**
  - We could improve a precision on the previous example... *data* → *positive real values...*

Numerical Data

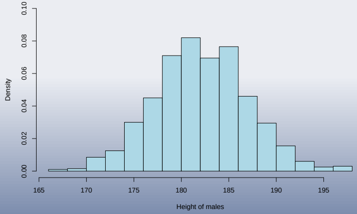
- **Real values...**
  - We could improve a precision on the previous example... *data* → *positive real values...*
  - How long can a machine work until it breaks down?

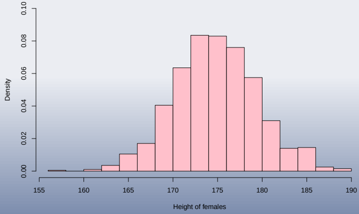
Numerical Data

- **Real values...**
  - We could improve a precision on the previous example... *data* $\rightarrow$ *positive real values...*
  - How long can a machine work until it breaks down?
- Data: a sequence of real valued observations...
  - infinitely many different outcomes to be considered;
  - the most common data presentation using histograms;

Numerical Data

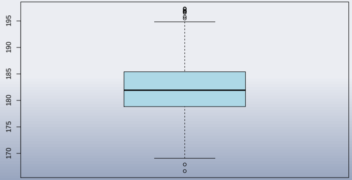
- **Real values...**
  - We could improve a precision on the previous example... *data* → *positive real values...*
  - How long can a machine work until it breaks down?
- Data: a sequence of real valued observations...
  - infinitely many different outcomes to be considered;
  - the most common data presentation using histograms;

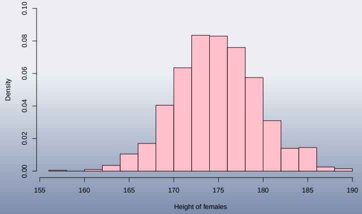




Numerical Data

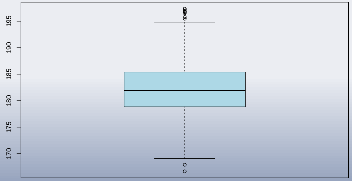
- **Real values...**
  - We could improve a precision on the previous example...
    data → positive real values...
  - How long can a machine work until it breaks down?
- Data: a sequence of real valued observations...
  - infinitely many different outcomes to be considered;
  - the most common data presentation using histograms;

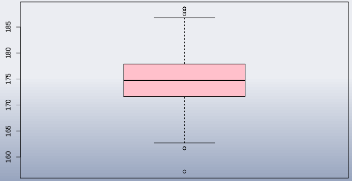




Numerical Data

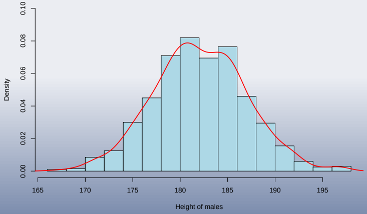
- **Real values...**
  - We could improve a precision on the previous example... *data* → *positive real values...*
  - How long can a machine work until it breaks down?
- Data: a sequence of real valued observations...
  - infinitely many different outcomes to be considered;
  - the most common data presentation using histograms;

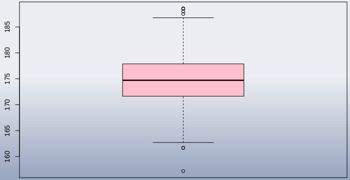




Numerical Data

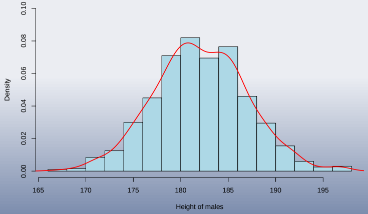
- **Real values...**
  - We could improve a precision on the previous example... *data → positive real values...*
  - How long can a machine work until it breaks down?
- Data: a sequence of real valued observations...
  - infinitely many different outcomes to be considered;
  - the most common data presentation using histograms;

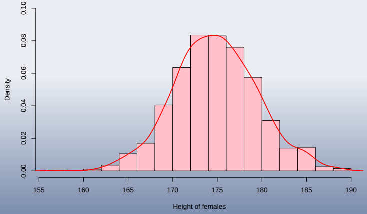




Numerical Data

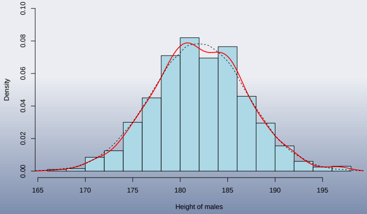
- **Real values...**
  - We could improve a precision on the previous example... *data → positive real values...*
  - How long can a machine work until it breaks down?
- Data: a sequence of real valued observations...
  - infinitely many different outcomes to be considered;
  - the most common data presentation using histograms;

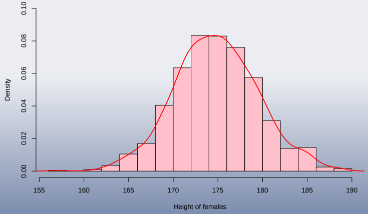




Numerical Data

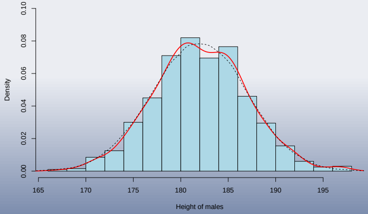
- **Real values...**
  - We could improve a precision on the previous example... *data → positive real values...*
  - How long can a machine work until it breaks down?
- Data: a sequence of real valued observations...
  - infinitely many different outcomes to be considered;
  - the most common data presentation using histograms;

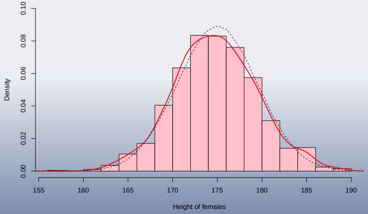




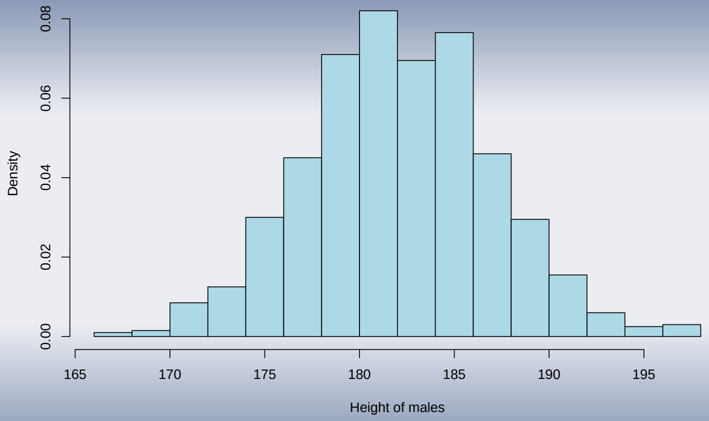
Numerical Data

- **Real values...**
  - We could improve a precision on the previous example... *data → positive real values...*
  - How long can a machine work until it breaks down?
- Data: a sequence of real valued observations...
  - infinitely many different outcomes to be considered;
  - the most common data presentation using histograms;

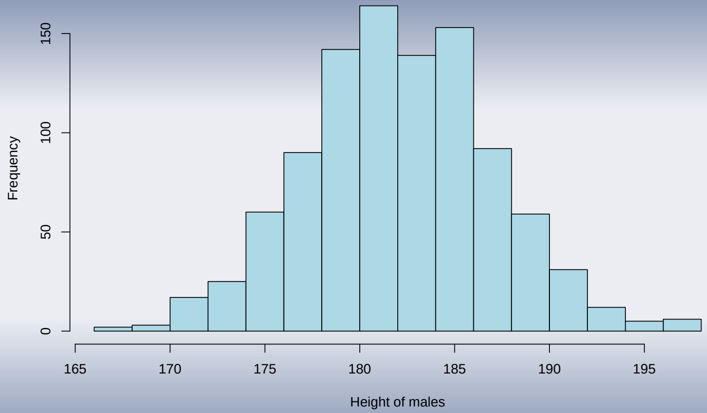




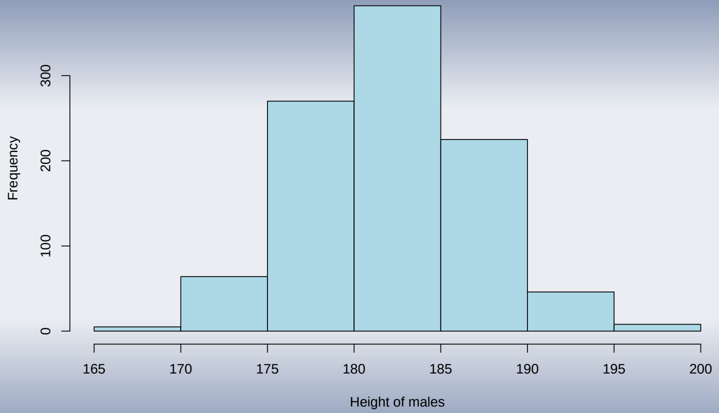
Histogram...


Density
0.00
0.02
0.04
0.06
0.08
165
170
175
180
185
190
195
Height of males

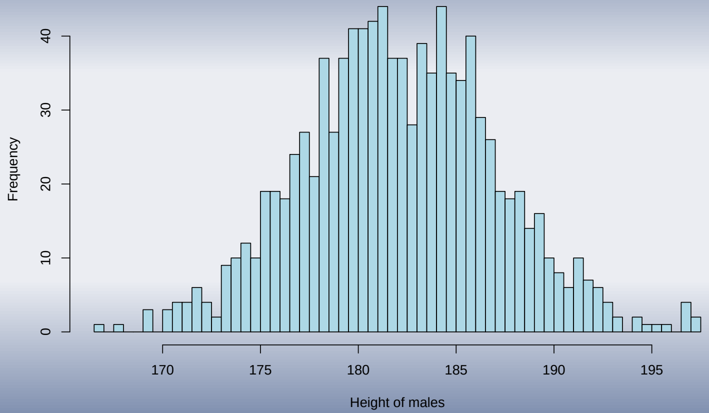
Histogram...


Frequency
0
50
100
150
165
170
175
180
185
190
195
Height of males

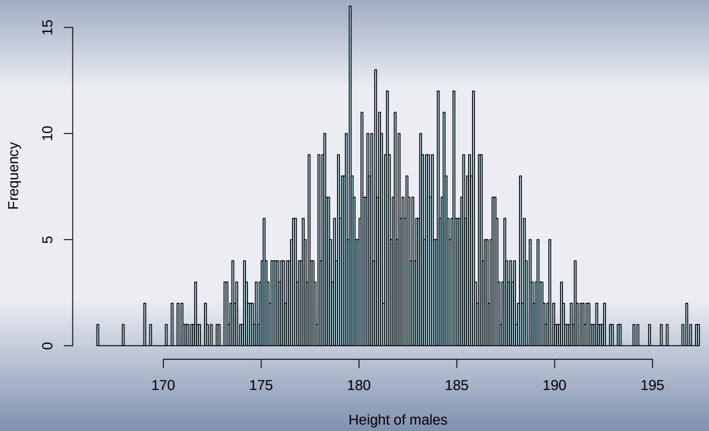
Histogram...


Frequency
0
100
200
300
165
170
175
180
185
190
195
200
Height of males

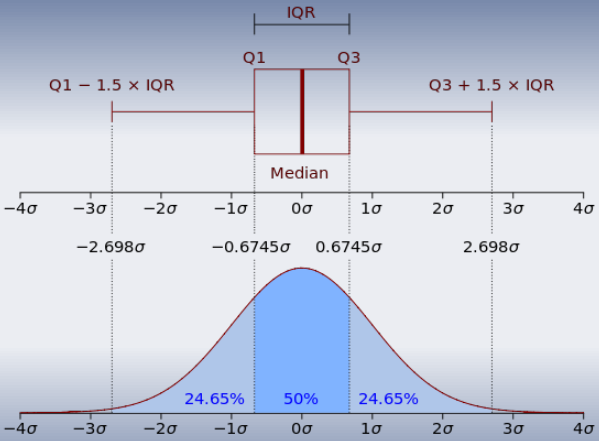
Histogram...



Histogram...



Boxplot figure...



Some nice properties

What can be directly observed from histogram (boxplot)?

❏ symmetric/non-symmetric data distribution;

❏ unimodal or multimodal data distribution;

❏ skewness of data distribution;

❏ indications of outlying observations;

❏ ...

Some nice properties

What can be directly observed from histogram (boxplot)?

- ❑ symmetric/non-symmetric data distribution;
- ❑ unimodal or multimodal data distribution;
- ❑ skewness of data distribution;
- ❑ indications of outlying observations;
- ❑ ...

It is not so much important now when we only focus on data exploration part however, it will become really crucial when considering more sophisticated statistical modeling approaches.

Data sample reliability

❏ How reliable such data samples are?

❏ Could we expect more reliability and confidence if more data is available?

Data sample reliability

- ❑ How reliable such data samples are?
- ❑ Could we expect more reliability and confidence if more data is available?

- ❑ Is there a way to numerically express such reliability, confidence?
- ❑ What should be a sufficient number of observation?

Data sample reliability

- ❏ How reliable such data samples are?
- ❏ Could we expect more reliability and confidence if more data is available?

- ❏ Is there a way to numerically express such reliability, confidence?
- ❏ What should be a sufficient number of observation?

- ❏ The more data we obtain the more we have to pay for...
- ❏ How to find a reasonable balance?

From counts to ordinal categories

- **How long does the machine work until it breaks down?**

From counts to ordinal categories

- **How long does the machine work until it breaks down?**
  - more that 15 hours ⇒ category A

From counts to ordinal categories

- **How long does the machine work until it breaks down?**
  - more that 15 hours ⇒ category A
  - more that 10 & less than 15 hours ⇒ category B

From counts to ordinal categories

- **How long does the machine work until it breaks down?**
  - more that 15 hours ⇒ category A
  - more that 10 & less than 15 hours ⇒ category B
  - more that 5 & less than 10 hours ⇒ category C

From counts to ordinal categories

- **How long does the machine work until it breaks down?**
  - more that 15 hours ⇒ category A
  - more that 10 & less than 15 hours ⇒ category B
  - more that 5 & less than 10 hours ⇒ category C
  - less than 5 hours ⇒ category D

From counts to ordinal categories

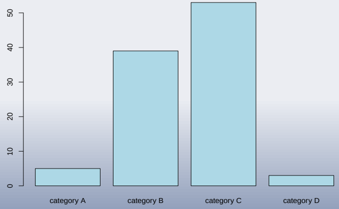
- **How long does the machine work until it breaks down?**
  - more that 15 hours ⇒ category A
  - more that 10 & less than 15 hours ⇒ category B
  - more that 5 & less than 10 hours ⇒ category C
  - less than 5 hours ⇒ category D

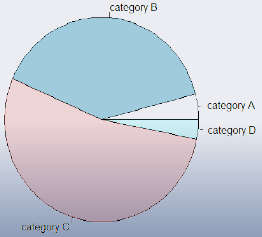
- **What is the data sample now?**

From counts to ordinal categories

- **How long does the machine work until it breaks down?**
  - more that 15 hours ⇒ category A
  - more that 10 & less than 15 hours ⇒ category B
  - more that 5 & less than 10 hours ⇒ category C
  - less than 5 hours ⇒ category D

- **What is the data sample now?**





Higher dimensional data

- two categorical variables;
- one categorical and the other numerical variable;
- two numerical (continuous) variables;

Higher dimensional data

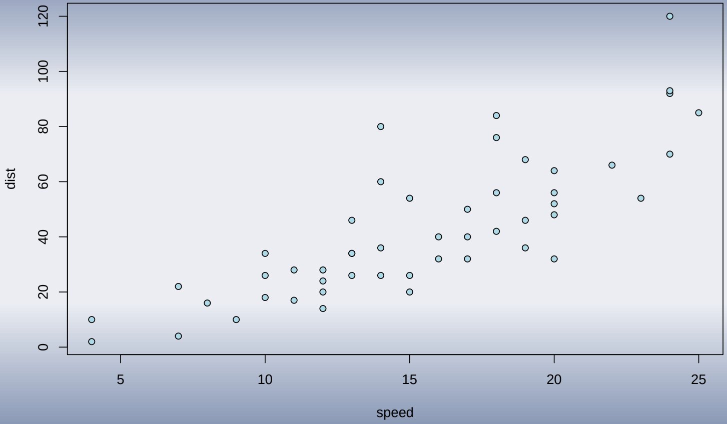
- two categorical variables;
- one categorical and the other numerical variable;
- two numerical (continuous) variables;

- more dimensional data $\Rightarrow$ it gets even difficult to plot;
- multivariate statistical methods were proposed to be used instead;

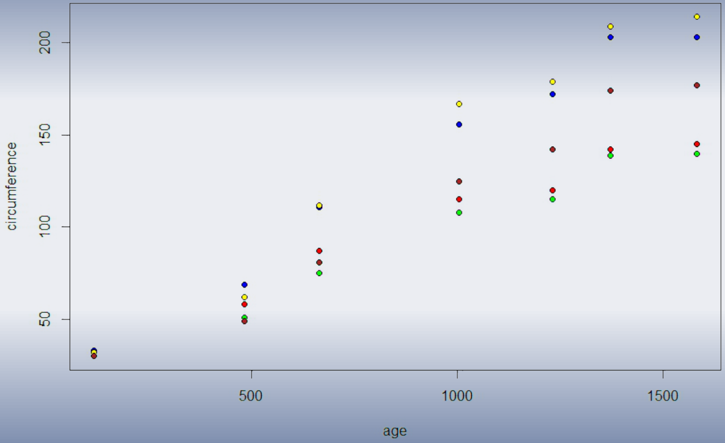
Titanic survivals

```
      Sex                                     Sex
Class  Male Female                  Class  Male Female
  1st     0      0                    1st     5      1
  2nd     0      0                    2nd    11     13
  3rd    35     17                    3rd    13     14
  Crew    0      0                    Crew    0      0

,, Age = Child, Survived = No       ,, Age = Child, Survived = Yes



      Sex                                     Sex
Class  Male Female                  Class  Male Female
  1st   118      4                    1st    57    140
  2nd   154     13                    2nd    14     80
  3rd   387     89                    3rd    75     76
  Crew  670      3                    Crew  192     20

,, Age = Adult, Survived = No       ,, Age = Adult, Survived = Yes
```

Car Breaking Distance



Orange Trees



From Population to its Sample

population $\Longrightarrow$ population sample $\Longrightarrow$ statistical inference

From Population to its Sample

population $\Longrightarrow$ population sample $\Longrightarrow$ statistical inference

Sampling procedure:

- ❑ ideally we would like to obtain a random sample;

From Population to its Sample

population $\Longrightarrow$ population sample $\Longrightarrow$ statistical inference

Sampling procedure:

- ❑ ideally we would like to obtain a random sample;
- ❑ random sample → independent and identically distributed observations;

Sample Statistics

Two different worlds

...

Sample Mean

- sample mean $\equiv$ average

Sample Mean

- sample mean $\equiv$ average $\neq$ mean

Sample Mean

- sample mean $\equiv$ average $\neq$ mean

- arithmetic average of all given observations;
- notation: $\overline{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$, where the actual observations are $\{x_1, x_2, \ldots, x_n\}$;

Sample Mean

❑ sample mean $\equiv$ average $\neq$ mean

- ❑ arithmetic average of all given observations;
- ❑ notation: $\overline{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$, where the actual observations are $\{x_1, x_2, \ldots, x_n\}$;
- ❑ it is a "middle value" with respect to values and their distribution;

Sample Mean

❑ sample mean $\equiv$ average $\neq$ mean

- ❑ arithmetic average of all given observations;
- ❑ notation: $\overline{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$, where the actual observations are $\{x_1, x_2, \ldots, x_n\}$;
- ❑ it is a "middle value" with respect to values and their distribution;

Sample Mean

❑ sample mean $\equiv$ average $\neq$ mean

- ❑ arithmetic average of all given observations;
- ❑ notation: $\overline{x}_n = \frac{\sum_{i=1}^n x_i}{n}$, where the actual observations are $\{x_1, x_2, \ldots, x_n\}$;
- ❑ it is a "middle value" with respect to values and their distribution;

- ❑ Advantage: it is sensitive with respect to outlying observations;

Sample Mean

❑ sample mean $\equiv$ average $\neq$ mean

- ❑ arithmetic average of all given observations;
- ❑ notation: $\overline{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$, where the actual observations are $\{x_1, x_2, \ldots, x_n\}$;
- ❑ it is a “middle value” with respect to values and their distribution;

- ❑ Advantage: it is sensitive with respect to outlying observations;
- ❑ Disadvantage: it is sensitive with respect to outlying observations;

Sample Mean

❑ sample mean $\equiv$ average $\neq$ mean

- ❑ arithmetic average of all given observations;
- ❑ notation: $\overline{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$, where the actual observations are $\{x_1, x_2, \dots, x_n\}$;
- ❑ it is a "middle value" with respect to values and their distribution;

- ❑ Advantage: it is sensitive with respect to outlying observations;
- ❑ Disadvantage: it is sensitive with respect to outlying observations;

- ❑ Some other proposals: sample trimmed mean, weighted mean, etc.;

Sample Median

- middle observation from all ordered observations;
- $\tilde{x}_n = x_{(\frac{n+1}{2})}$ for n even and $\tilde{x}_n = [x_{(n/2)} + x_{(\frac{n+1}{2})}]/2$ for n odd;

Sample Median

- middle observation from all ordered observations;
- $\tilde{x}_n = x_{(\frac{n+1}{2})}$ for n even and $\tilde{x}_n = [x_{(n/2)} + x_{(\frac{n+1}{2})}]/2$ for n odd;
- it is a "middle value" with respect to given values themselves;

Sample Median

- ❑ middle observation from all ordered observations;
- ❑ $\tilde{x}_n = x_{(\frac{n+1}{2})}$ for n even and $\tilde{x}_n = [x_{(n/2)} + x_{(\frac{n+1}{2})}]/2$ for n odd;
- ❑ it is a "middle value" with respect to given values themselves;

Sample Median

- ❑ middle observation from all ordered observations;
- ❑ $\tilde{x}_n = x_{(\frac{n+1}{2})}$ for n even and $\tilde{x}_n = [x_{(n/2)} + x_{(\frac{n+1}{2})}]/2$ for n odd;
- ❑ it is a "middle value" with respect to given values themselves;

- ❑ Advantage: it is not sensitive with respect to observation values;

Sample Median

- ❑ middle observation from all ordered observations;
- ❑ $\tilde{x}_n = x_{(\frac{n+1}{2})}$ for n even and $\tilde{x}_n = [x_{(n/2)} + x_{(\frac{n+1}{2})}]/2$ for n odd;
- ❑ it is a "middle value" with respect to given values themselves;

- ❑ Advantage: it is not sensitive with respect to observation values;
- ❑ Disadvantage: it is not sensitive with respect to observations values;

Sample Median

- ❑ middle observation from all ordered observations;
- ❑ $\tilde{x}_n = x_{(\frac{n+1}{2})}$ for n even and $\tilde{x}_n = [x_{(n/2)} + x_{(\frac{n+1}{2})}]/2$ for n odd;
- ❑ it is a "middle value" with respect to given values themselves;

- ❑ Advantage: it is not sensitive with respect to observation values;
- ❑ Disadvantage: it is not sensitive with respect to observations values;

- ❑ Some other proposals as well ...;

Some other sample characteristics

- **sample mode (for categorical variables mostly);**
  What is the relation between sample mean, median and mode?
- **sample variance & sample standard error;**
  mostly it is difficult to imagine → *sample range used "instead"*
- **sample quantiles (quartiles, percentiles, etc.);**
  some of them are more important than others...
- **coefficient of variation;**
  spread of data relative to its middle value

To be continued...

- Probability, probability concepts;
- Random events, combination of events;
- Conditional probability;
- Independence principle;
- Law of Total Probability;
- Bayes Theorem in Probability;