STATISTICAL TERMINOLOGY

The Basics, Misconceptions, and Pedantises

AARHUS UNIVERSITY

Erik Kusch

erik.kusch@au.dk

Section for Ecoinformatics & Biodiversity
Center for Biodiversity and Dynamics in a Changing World (BIOCHANGE)
Aarhus University

04/03/2020

Population vs. Sample

Population: describes the sum total of all *existing* values of a variable given a certain research question. This includes non-measured data.

Sample: describes the sum total of all *available* values of a variable for any given analysis. This can only include measured data.

An example:

In an experimental set-up, you rear an ant colony of exactly 10,000 individuals. You are interested in the average mandible strength of ants within the colony.

The problem: You cannot possibly take measurements of all 10,000 individuals.

The solution: Taking measurements on a **Sample** (e.g. 1,000 individuals) from within the **Population** (10,000 individuals).

Population vs. Sample

Population: describes the sum total of all *existing* values of a variable given a certain research question. This includes non-measured data.

Sample: describes the sum total of all *available* values of a variable for any given analysis. This can only include measured data.

An example:

In an experimental set-up, you rear an ant colony of exactly 10,000 individuals. You are interested in the average mandible strength of ants within the colony.

The problem: You cannot possibly take measurements of all 10,000 individuals.

The solution: Taking measurements on a **Sample** (e.g. 1,000 individuals) from within the **Population** (10,000 individuals).

Population vs. Sample

Population: describes the sum total of all *existing* values of a variable given a certain research question. This includes non-measured data.

Sample: describes the sum total of all *available* values of a variable for any given analysis. This can only include measured data.

An example:

In an experimental set-up, you rear an ant colony of exactly 10,000 individuals. You are interested in the average mandible strength of ants within the colony.

The problem: You cannot possibly take measurements of all 10,000 individuals.

The solution: Taking measurements on a **Sample** (e.g. 1,000 individuals) from within the **Population** (10,000 individuals).

Population vs. Sample

Population: describes the sum total of all *existing* values of a variable given a certain research question. This includes non-measured data.

Sample: describes the sum total of all *available* values of a variable for any given analysis. This can only include measured data.

An example:

In an experimental set-up, you rear an ant colony of exactly 10,000 individuals. You are interested in the average mandible strength of ants within the colony.

The problem: You cannot possibly take measurements of all 10,000 individuals.

The solution: Taking measurements on a **Sample** (e.g. 1,000 individuals) from within the **Population** (10,000 individuals).

Population vs. Sample

Population: describes the sum total of all *existing* values of a variable given a certain research question. This includes non-measured data.

Sample: describes the sum total of all *available* values of a variable for any given analysis. This can only include measured data.

An example:

In an experimental set-up, you rear an ant colony of exactly 10,000 individuals. You are interested in the average mandible strength of ants within the colony.

The problem: You cannot possibly take measurements of all 10,000 individuals.

The solution: Taking measurements on a **Sample** (e.g. 1,000 individuals) from within the **Population** (10,000 individuals).

Population vs. Sample

Population: describes the sum total of all *existing* values of a variable given a certain research question. This includes non-measured data.

Sample: describes the sum total of all *available* values of a variable for any given analysis. This can only include measured data.

An example:

In an experimental set-up, you rear an ant colony of exactly 10,000 individuals. You are interested in the average mandible strength of ants within the colony.

The problem: You cannot possibly take measurements of all 10,000 individuals.

The solution: Taking measurements on a **Sample** (e.g. 1,000 individuals) from within the **Population** (10,000 individuals).

Test- vs. Training-Data

This differentiation is only applicable when concerned with *modelling*.

Training Data: describes the subset of the total data which is used to *establish/train* the model.

Test Data: describes the subset of the total data which is used to *test* the performance of the model.

The problem: You have identified a way to model how mandible strength and ant size are interconnected but don't know how to assess the quality of your model (a model will always fit the data it was built on extremely well).

The solution: Split the available data into two non-overlapping subsets of data (**Training** and **Test Data**) and use these separately to build your model and assess its performance.

Test- vs. Training-Data

This differentiation is only applicable when concerned with *modelling*.

Training Data: describes the subset of the total data which is used to *establish/train* the model.

Test Data: describes the subset of the total data which is used to *test* the performance of the model.

The problem: You have identified a way to model how mandible strength and ant size are interconnected but don't know how to assess the quality of your model (a model will always fit the data it was built on extremely well).

The solution: Split the available data into two non-overlapping subsets of data (**Training** and **Test Data**) and use these separately to build your model and assess its performance.

Test- vs. Training-Data

This differentiation is only applicable when concerned with *modelling*.

Training Data: describes the subset of the total data which is used to *establish/train* the model.

Test Data: describes the subset of the total data which is used to *test* the performance of the model.

The problem: You have identified a way to model how mandible strength and ant size are interconnected but don't know how to assess the quality of your model (a model will always fit the data it was built on extremely well).

The solution: Split the available data into two non-overlapping subsets of data (**Training** and **Test Data**) and use these separately to build your model and assess its performance.

Test- vs. Training-Data

This differentiation is only applicable when concerned with *modelling*.

Training Data: describes the subset of the total data which is used to *establish/train* the model.

Test Data: describes the subset of the total data which is used to *test* the performance of the model.

The problem: You have identified a way to model how mandible strength and ant size are interconnected but don't know how to assess the quality of your model (a model will always fit the data it was built on extremely well).

The solution: Split the available data into two non-overlapping subsets of data (**Training** and **Test Data**) and use these separately to build your model and assess its performance.

Test- vs. Training-Data

This differentiation is only applicable when concerned with *modelling*.

Training Data: describes the subset of the total data which is used to *establish/train* the model.

Test Data: describes the subset of the total data which is used to *test* the performance of the model.

The problem: You have identified a way to model how mandible strength and ant size are interconnected but don't know how to assess the quality of your model (a model will always fit the data it was built on extremely well).

The solution: Split the available data into two non-overlapping subsets of data (**Training** and **Test Data**) and use these separately to build your model and assess its performance.

Randomness

Randomisation is one of the **most important** practices in biological studies.

A **sampling** procedure is **random** when any member of the *population* has an equal chance of being selected into the *sample*.

Training and *Test Data Sets* are established from the population with the same sense of randomness although there may be exceptions depending on the modelling procedure at hand.

Data collection: Number all units contained within the set-up and sample those units corresponding to random numbers.

In R: Use the `sample()` function to create truly random subsets. Remember to use `set.seed()` to make this step reproducible!

Randomness

Randomisation is one of the **most important** practices in biological studies.

A **sampling** procedure is **random** when any member of the *population* has an equal chance of being selected into the *sample*.

Training and *Test Data Sets* are established from the population with the same sense of randomness although there may be exceptions depending on the modelling procedure at hand.

Data collection: Number all units contained within the set-up and sample those units corresponding to random numbers.

In R: Use the `sample()` function to create truly random subsets. Remember to use `set.seed()` to make this step reproducible!

Randomness

Randomisation is one of the **most important** practices in biological studies.

A **sampling** procedure is **random** when any member of the *population* has an equal chance of being selected into the *sample*.

Training and *Test Data Sets* are established from the population with the same sense of randomness although there may be exceptions depending on the modelling procedure at hand.

Data collection: Number all units contained within the set-up and sample those units corresponding to random numbers.

In R: Use the `sample()` function to create truly random subsets. Remember to use `set.seed()` to make this step reproducible!

Randomness

Randomisation is one of the **most important** practices in biological studies.

A **sampling** procedure is **random** when any member of the *population* has an equal chance of being selected into the *sample*.

Training and *Test Data Sets* are established from the population with the same sense of randomness although there may be exceptions depending on the modelling procedure at hand.

Data collection: Number all units contained within the set-up and sample those units corresponding to random numbers.

In R: Use the `sample()` function to create truly random subsets. Remember to use `set.seed()` to make this step reproducible!

Stratified Sampling

When do we break *true* **randomness**?.

When a **population** can be divided into distinct categories (i.e. **strata**). These can be regarded as individual sub-populations.

Stratified sampling ensures that all sub-populations are proportionally represented in the final population-sample given their relative size.

```
d
```

```
##   s Freq
## 1 A   50
## 2 B   35
## 3 C   15
```

```
set.seed(42)  # stratified
table(sample(d$s, replace = TRUE, prob = d$Freq, 100))
```

```
##
##  A  B  C
## 45 38 17
```

```
set.seed(42)  # non-stratified
table(sample(d$s, replace = TRUE, 100))
```

```
##
##  A  B  C
## 40 39 21
```

Stratified Sampling

When do we break *true* **randomness**?.

When a **population** can be divided into distinct categories (i.e. **strata**). These can be regarded as individual sub-populations.

Stratified sampling ensures that all sub-populations are proportionally represented in the final population-sample given their relative size.

```
d
```

```
##   s Freq
## 1 A   50
## 2 B   35
## 3 C   15
```

```
set.seed(42)  # stratified
table(sample(d$s, replace = TRUE, prob = d$Freq, 100))
```

```
##
##  A  B  C
## 45 38 17
```

```
set.seed(42)  # non-stratified
table(sample(d$s, replace = TRUE, 100))
```

```
##
##  A  B  C
## 40 39 21
```

Stratified Sampling

When do we break *true* **randomness**?.

When a **population** can be divided into distinct categories (i.e. **strata**). These can be regarded as individual sub-populations.

Stratified sampling ensures that all sub-populations are proportionally represented in the final population-sample given their relative size.

```r
d
```

```
##   s Freq
## 1 A   50
## 2 B   35
## 3 C   15
```

```r
set.seed(42)  # stratified
table(sample(d$s, replace = TRUE, prob = d$Freq, 100))
```

```
##
##  A  B  C
## 45 38 17
```

```r
set.seed(42)  # non-stratified
table(sample(d$s, replace = TRUE, 100))
```

```
##
##  A  B  C
## 40 39 21
```

Unsupervised Approaches

Unsupervised methods are often *used to select* the *most informative X input variables* for *supervised approaches*.

Pre-requisites:

- Only *input variables* are observed.
- *No solution/feedback (output)* is given.

Aims:

- *Divide* the observations into relatively distinct groups.
- *Model* the underlying structure or distribution in the data.

$\rightarrow$ **"Pre-processing" before a supervised learning analysis** and **exploratory analyses**

Unsupervised Approaches

Unsupervised methods are often *used to select* the *most informative* X *input variables* for *supervised approaches*.

Pre-requisites:

- Only *input variables* are observed.
- *No solution/feedback (output)* is given.

Aims:

- *Divide* the observations into relatively distinct groups.
- *Model* the underlying structure or distribution in the data.

$\rightarrow$ **"Pre-processing" before a supervised learning analysis** and **exploratory analyses**

Unsupervised Approaches

Unsupervised methods are often *used to select* the *most informative* X *input variables* for *supervised approaches*.

Pre-requisites:

- Only *input variables* are observed.
- *No solution/feedback (output)* is given.

Aims:

- *Divide* the observations into relatively distinct groups.
- *Model* the underlying structure or distribution in the data.

$\rightarrow$ **"Pre-processing" before a supervised learning analysis** and **exploratory analyses**

Unsupervised Approaches

Unsupervised methods are often *used to select* the *most informative X input variables* for *supervised approaches*.

Pre-requisites:

- Only *input variables* are observed.
- *No solution/feedback (output)* is given.

Aims:

- *Divide* the observations into relatively distinct groups.
- *Model* the underlying structure or distribution in the data.

$\rightarrow$ **"Pre-processing" before a supervised learning analysis** and **exploratory analyses**

Supervised Approaches

Supervised methods are often *informed by unsupervised approaches* and used to *gain validated information* about the data.

Pre-requisites:

- Both *predictors* $\mathbf{X}$, and *responses* $\mathbf{Y}$ are observed (there is one y_i for each x_i).
- Data is split into *Training* and *Test Data Sets*.

Aims:

- Learn a *mapping function* f from X to Y.
- *Validate* established function/model.
- Further *prediction* and *inference*.

$\rightarrow$ **Mostly inferential analyses**

Supervised Approaches

Supervised methods are often *informed by unsupervised approaches* and used to *gain validated information* about the data.

Pre-requisites:

- Both *predictors* $\mathbf{X}$, and *responses* $\mathbf{Y}$ are observed (there is one y_i for each x_i).
- Data is split into *Training* and *Test Data Sets*.

Aims:

- Learn a *mapping function* f from X to Y.
- *Validate* established function/model.
- Further *prediction* and *inference*.

$\rightarrow$ **Mostly inferential analyses**

Supervised Approaches

Supervised methods are often *informed by unsupervised approaches* and used to *gain validated information* about the data.

Pre-requisites:

- Both *predictors* $\mathbf{X}$, and *responses* $\mathbf{Y}$ are observed (there is one y_i for each x_i).
- Data is split into *Training* and *Test Data Sets*.

Aims:

- Learn a *mapping function* f from X to Y.
- *Validate* established function/model.
- Further *prediction* and *inference*.

$\rightarrow$ **Mostly inferential analyses**

Supervised Approaches

Supervised methods are often *informed by unsupervised approaches* and used to *gain validated information* about the data.

Pre-requisites:

- Both *predictors* $\mathbf{X}$, and *responses* $\mathbf{Y}$ are observed (there is one y_i for each x_i).
- Data is split into *Training* and *Test Data Sets*.

Aims:

- Learn a *mapping function* f from X to Y.
- *Validate* established function/model.
- Further *prediction* and *inference*.

$\rightarrow$ **Mostly inferential analyses**

Types of Variables

Variables can be classed into a multitude of types. The most common classification system knows:

Categorical Variables

- also known as *Qualitative Variables*
- Scales can be either:
  - Nominal
  - Ordinal

Continuous Variables

- also known as *Quantitative Variables*
- Scales can be either:
  - Discrete
  - Continuous

Types of Variables

Variables can be classed into a multitude of types. The most common classification system knows:

Categorical Variables

- also known as *Qualitative Variables*
- Scales can be either:
  - Nominal
  - Ordinal

Continuous Variables

- also known as *Quantitative Variables*
- Scales can be either:
  - Discrete
  - Continuous

Types of Variables

Variables can be classed into a multitude of types. The most common classification system knows:

Categorical Variables

- also known as *Qualitative Variables*
- Scales can be either:
  - Nominal
  - Ordinal

Continuous Variables

- also known as *Quantitative Variables*
- Scales can be either:
  - Discrete
  - Continuous

Categorical Variables

Categorical variables are those variables which **establish and fall into distinct groups and classes**.

Categorical variables:

- can take on a finite number of values
- assign each unit of the population to one of a finite number of groups
- can *sometimes* be ordered

In R, categorical variables usually come up as object type `factor` or `character`.

Categorical Variables

Categorical variables are those variables which **establish and fall into distinct groups and classes**.

Categorical variables:

- can take on a finite number of values
- assign each unit of the population to one of a finite number of groups
- can *sometimes* be ordered

In R, categorical variables usually come up as object type `factor` or `character`.

Categorical Variables

Categorical variables are those variables which **establish and fall into distinct groups and classes**.

Categorical variables:

- can take on a finite number of values
- assign each unit of the population to one of a finite number of groups
- can *sometimes* be ordered

In R, categorical variables usually come up as object type `factor` or `character`.

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Categorical Variables (Examples)

Examples of categorical variables:

- Biome Classifications (e.g. "Boreal Forest", "Tundra", etc.)
- Sex (e.g. "Male", "Female")
- Hierarchy Position (e.g. "α-Individual", "β-Individual", etc.)
- Soil Type (e.g. "Sandy", "Mud", "Permafrost", etc.)
- Leaf Type (e.g. "Compound", "Single Blade", etc.)
- Sexual Reproductive Stage (e.g. "Juvenile", "Mature", etc.)
- Species Membership
- Family Group Membership
- ...

Continuous Variables

Continuous variables are those variables which **establish a range of possible data values**.

Continuous variables:

- can take on an infinite number of values
- can take on a new value for each unit in the set-up
- can *always* be ordered

In R, continuous variables usually come up as object type `numeric`.

Continuous Variables

Continuous variables are those variables which **establish a range of possible data values**.

Continuous variables:

- can take on an infinite number of values
- can take on a new value for each unit in the set-up
- can *always* be ordered

In R, continuous variables usually come up as object type `numeric`.

Continuous Variables

Continuous variables are those variables which **establish a range of possible data values**.

Continuous variables:

- can take on an infinite number of values
- can take on a new value for each unit in the set-up
- can *always* be ordered

In R, continuous variables usually come up as object type `numeric`.

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Continuous Variables (Examples)

Examples of categorical variables:

- Temperature
- Precipitation
- Weight
- pH
- Altitude
- Group Size
- Vegetation Indices
- Time
- ...

Converting Variable Types

Continuous variables can be converted into *categorical variables* via a method called **binning:**

Given a variable range, one can establish however many "bins" as one wants. For example:

- Given a temperature range of $271K - 291K$, there may be 4 bins of equal size:
  - Bin A: $271K \leq X \leq 276K$
  - Bin B: $276K < X \leq 281K$
  - Bin C: $281K < X \leq 286K$
  - Bin D: $286K < X \leq 291K$

Whilst a **continuous variable** can be both *continuous* and *categorical*, a **categorical variable** can only ever be *categorical*!

Converting Variable Types

Continuous variables can be converted into *categorical variables* via a method called **binning:**

Given a variable range, one can establish however many “bins” as one wants. For example:

- Given a temperature range of $271K - 291K$, there may be 4 bins of equal size:
  - Bin A: $271K \leq X \leq 276K$
  - Bin B: $276K < X \leq 281K$
  - Bin C: $281K < X \leq 286K$
  - Bin D: $286K < X \leq 291K$

Whilst a **continuous variable** can be both *continuous* and *categorical*, a **categorical variable** can only ever be *categorical*!

Converting Variable Types

Continuous variables can be converted into *categorical variables* via a method called **binning:**

Given a variable range, one can establish however many "bins" as one wants. For example:

- Given a temperature range of $271K - 291K$, there may be 4 bins of equal size:
  - Bin A: $271K \leq X \leq 276K$
  - Bin B: $276K < X \leq 281K$
  - Bin C: $281K < X \leq 286K$
  - Bin D: $286K < X \leq 291K$

Whilst a **continuous variable** can be both *continuous* and *categorical*, a **categorical variable** can only ever be *categorical*!

Converting Variable Types

Continuous variables can be converted into *categorical variables* via a method called **binning:**

Given a variable range, one can establish however many "bins" as one wants.

For example:

- Given a temperature range of $271K - 291K$, there may be 4 bins of equal size:
  - Bin A: $271K \leq X \leq 276K$
  - Bin B: $276K < X \leq 281K$
  - Bin C: $281K < X \leq 286K$
  - Bin D: $286K < X \leq 291K$

Whilst a **continuous variable** can be both *continuous* and *categorical*, a **categorical variable** can only ever be *categorical*!

Variables On Scales

Another way of classifying variables are the **scales** they are represented on.

Different scales of variables **require different statistical procedures** for analyses!

Variable scales include:

- **Nominal**
- **Binary**
- **Ordinal**
- **Interval**
- **Relation/Ratio**

Some statistics books teach *integer scales* along the above mentioned scales. Some people dispute this and claim these scales to be *ratio scales*.

Variables On Scales

Another way of classifying variables are the **scales** they are represented on.

Different scales of variables **require different statistical procedures** for analyses!

Variable scales include:

- **Nominal**
- **Binary**
- **Ordinal**
- **Interval**
- **Relation/Ratio**

Some statistics books teach *integer scales* along the above mentioned scales. Some people dispute this and claim these scales to be *ratio scales*.

Variables On Scales

Another way of classifying variables are the **scales** they are represented on.

Different scales of variables **require different statistical procedures** for analyses!

Variable scales include:

- **Nominal**
- **Binary**
- **Ordinal**
- **Interval**
- **Relation/Ratio**

Some statistics books teach *integer scales* along the above mentioned scales. Some people dispute this and claim these scales to be *ratio scales*.

Variables On Scales

Another way of classifying variables are the **scales** they are represented on.

Different scales of variables **require different statistical procedures** for analyses!

Variable scales include:

- **Nominal**
- **Binary**
- **Ordinal**
- **Interval**
- **Relation/Ratio**

Some statistics books teach *integer scales* along the above mentioned scales. Some people dispute this and claim these scales to be *ratio scales*.

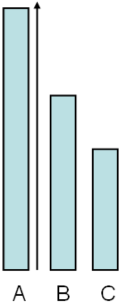
Nominal And Binary

Nominal scales of variables correspond to *categorical variables* which cannot be put into a meaningful order.

- Variables on nominal scales put units into distinct categories
- These variables may be numerical but offer no mathematical interpretation

Examples:

- Petal colour (red, green, blue, etc.)
- Individual IDs

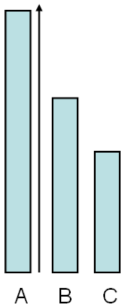


Binary scales are a special case of *nominal scales* taking only two possible values: 0 and 1.

Nominal And Binary

Nominal scales of variables correspond to *categorical variables* which cannot be put into a meaningful order.

- Variables on nominal scales put units into distinct categories
- These variables may be numerical but offer no mathematical interpretation



Examples:

- Petal colour (red, green, blue, etc.)
- Individual IDs

Binary scales are a special case of *nominal scales* taking only two possible values: 0 and 1.

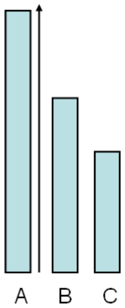
Nominal And Binary

Nominal scales of variables correspond to *categorical variables* which cannot be put into a meaningful order.

- Variables on nominal scales put units into distinct categories
- These variables may be numerical but offer no mathematical interpretation

Examples:

- Petal colour (red, green, blue, etc.)
- Individual IDs



Binary scales are a special case of *nominal scales* taking only two possible values: 0 and 1.

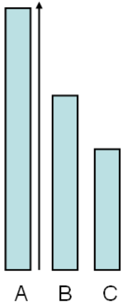
Nominal And Binary

Nominal scales of variables correspond to *categorical variables* which cannot be put into a meaningful order.

- Variables on nominal scales put units into distinct categories
- These variables may be numerical but offer no mathematical interpretation

Examples:

- Petal colour (red, green, blue, etc.)
- Individual IDs



Binary scales are a special case of *nominal scales* taking only two possible values: 0 and 1.

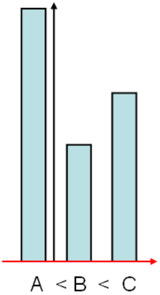
Ordinal

Ordinal scales of variables correspond to *categorical variables* which can be put into meaningful order.

- Variables on ordinal scales put units into distinct categories
- These variables may be numerical and offer some mathematical interpretation

Examples:

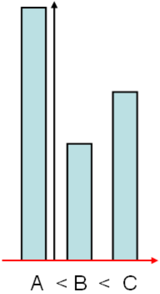
- Size (small, medium, large, etc.)
- Binned continuous variables



Ordinal

Ordinal scales of variables correspond to *categorical variables* which can be put into meaningful order.

- Variables on ordinal scales put units into distinct categories
- These variables may be numerical and offer some mathematical interpretation



Examples:

- Size (small, medium, large, etc.)
- Binned continuous variables

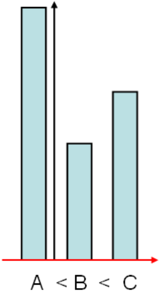
Ordinal

Ordinal scales of variables correspond to *categorical variables* which can be put into meaningful order.

- Variables on ordinal scales put units into distinct categories
- These variables may be numerical and offer some mathematical interpretation

Examples:

- Size (small, medium, large, etc.)
- Binned continuous variables



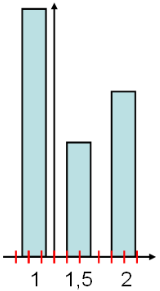
Interval/Discrete

Interval scales of variables correspond to a mix of *continuous variables*.

- Variables on interval scales are measured on equal intervals from a defined zero point/point of origin
- The point of origin **does not imply an absence of the measured characteristic**

Examples:

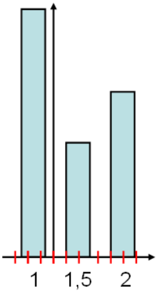
- Temperature [℃]
- pH



Interval/Discrete

Interval scales of variables correspond to a mix of *continuous variables*.

- Variables on interval scales are measured on equal intervals from a defined zero point/point of origin
- The point of origin **does not imply an absence of the measured characteristic**



Examples:

- Temperature [℃]
- pH

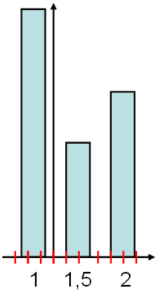
Interval/Discrete

Interval scales of variables correspond to a mix of *continuous variables*.

- Variables on interval scales are measured on equal intervals from a defined zero point/point of origin
- The point of origin **does not imply an absence of the measured characteristic**

Examples:

- Temperature [℃]
- pH



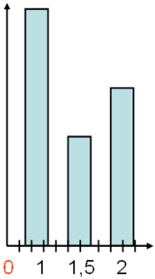
Relation/Ratio

Relation/Ratio scales of variables correspond to *continuous variables*.

- Variables on relation/ratio scales are measured on equal intervals from a defined zero point/point of origin
- The point of origin **does imply an absence of the measured characteristic**

Examples:

- Temperature $[K]$
- Weight

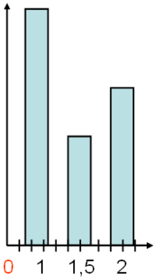


Integer scales are a special case of *ratio scales* allowing only for integral numbers.

Relation/Ratio

Relation/Ratio scales of variables correspond to *continuous variables*.

- Variables on relation/ratio scales are measured on equal intervals from a defined zero point/point of origin
- The point of origin **does imply an absence of the measured characteristic**



Examples:

- Temperature [K]
- Weight

Integer scales are a special case of *ratio scales* allowing only for integral numbers.

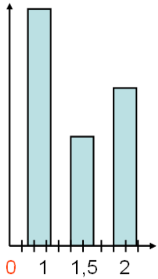
Relation/Ratio

Relation/Ratio scales of variables correspond to *continuous variables*.

- Variables on relation/ratio scales are measured on equal intervals from a defined zero point/point of origin
- The point of origin **does imply an absence of the measured characteristic**

Examples:

- Temperature $[K]$
- Weight



Integer scales are a special case of *ratio scales* allowing only for integral numbers.

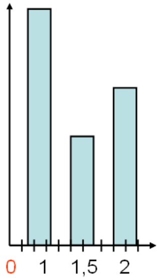
Relation/Ratio

Relation/Ratio scales of variables correspond to *continuous variables*.

- Variables on relation/ratio scales are measured on equal intervals from a defined zero point/point of origin
- The point of origin **does imply an absence of the measured characteristic**

Examples:

- Temperature $[K]$
- Weight



Integer scales are a special case of *ratio scales* allowing only for integral numbers.

Confusion Of Units

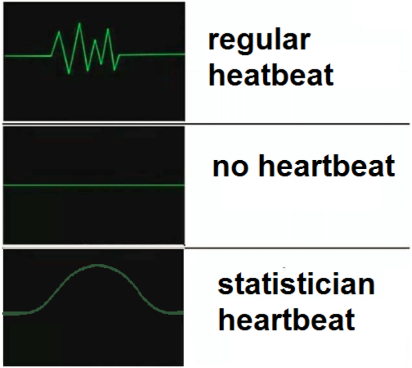
What Are Distributions?

A distribution of a statistical data set (sample/population) shows all the possible values/intervals of the data in question and their frequency.

$\rightarrow$ Basically, **data patterns** we are considering/looking for.

What Are Distributions?

A distribution of a statistical data set (sample/population) shows all the possible values/intervals of the data in question and their frequency.

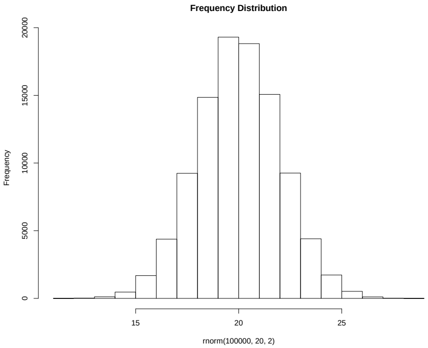


→ Basically, **data patterns** we are considering/looking for.

Frequency Distributions

Frequency Distributions:

- *Theory*
  - Simple representations of data value frequencies
  - Can be established for every variable
- *Practice in R*
  - Visualisation via the ‘hist()‘ function

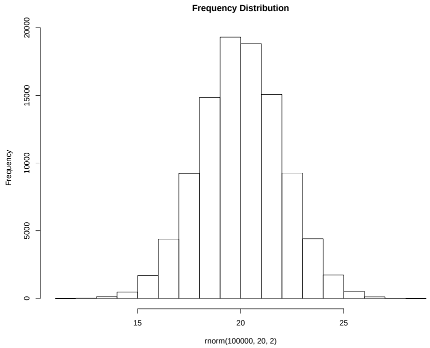



```
hist(rnorm(100000,20,2),
main = "Frequency Distribution")
```

Frequency Distributions

Frequency Distributions:

- *Theory*
  - Simple representations of data value frequencies
  - Can be established for every variable
- *Practice in R*
  - Visualisation via the ‘hist()‘ function

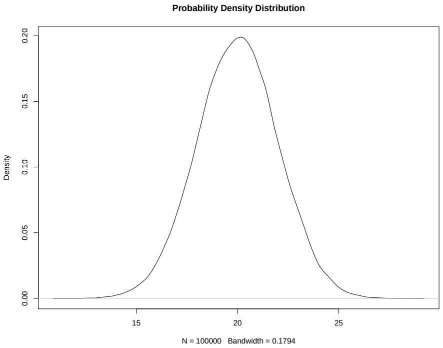



```
hist(rnorm(100000,20,2),
main = "Frequency Distribution")
```

Probability Density Distributions I

Probability Density Distributions:

- *Theory*
  - Representation of data value probabilities
  - Can be established for *continuous* variables
- *Practice in R*
  - Visualisation via the 'density()' function

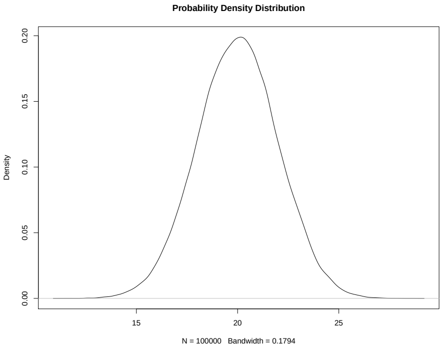



```
plot(density(rnorm(100000,20,2)),
  main = "Probability Density Distribution")
```

Probability Density Distributions I

Probability Density Distributions:

- *Theory*
  - Representation of data value probabilities
  - Can be established for *continuous* variables
- *Practice in R*
  - Visualisation via the ‘density()‘ function




```
plot(density(rnorm(100000,20,2)),
  main = "Probability Density Distribution")
```

Probability Density Distributions II

Probability Density Distributions hold the **majority of importance** in statistics!

A few key points about these distributions:

- Area under the curve (AUC) sums to 1
- A probability for every given single value is 0
- The AUC between two values on the X-axis equals the probability to randomly sample a value between these two points

Find a masterful explanation of the single-value probability here.

Probability Density Distributions II

Probability Density Distributions hold the **majority of importance** in statistics!

A few key points about these distributions:

- Area under the curve (AUC) sums to 1
- A probability for every given single value is 0
- The AUC between two values on the X-axis equals the probability to randomly sample a value between these two points

Find a masterful explanation of the single-value probability here.

Probability Density Distributions II

Probability Density Distributions hold the **majority of importance** in statistics!

A few key points about these distributions:

- Area under the curve (AUC) sums to 1
- A probability for every given single value is 0
- The AUC between two values on the X-axis equals the probability to randomly sample a value between these two points

Find a masterful explanation of the single-value probability here.

Probability Density Distributions II

Probability Density Distributions hold the **majority of importance** in statistics!

A few key points about these distributions:

- Area under the curve (AUC) sums to 1
- A probability for every given single value is 0
- The AUC between two values on the X-axis equals the probability to randomly sample a value between these two points

Find a masterful explanation of the single-value probability here.

Probability Density Distributions II

Probability Density Distributions hold the **majority of importance** in statistics!

A few key points about these distributions:

- Area under the curve (AUC) sums to 1
- A probability for every given single value is 0
- The AUC between two values on the X-axis equals the probability to randomly sample a value between these two points

Find a masterful explanation of the single-value probability here.

Probability Density Distributions II

Probability Density Distributions hold the **majority of importance** in statistics!

A few key points about these distributions:

- Area under the curve (AUC) sums to 1
- A probability for every given single value is 0
- The AUC between two values on the X-axis equals the probability to randomly sample a value between these two points

Find a masterful explanation of the single-value probability here.

Univariate Standard Normal/Gaussian Distribution

One of the **most important** distributions in natural sciences.

- Used to represent real-valued random variables whose distributions are not known
- The **central limit theorem** applies (draw a sufficient number of samples and you end up with the normal distribution)
- These distributions are usually known also as "bell curves" (**Attention:** other distributions take this shape too)

Univariate Standard Normal/Gaussian Distribution

One of the **most important** distributions in natural sciences.

- Used to represent real-valued random variables whose distributions are not known
- The **central limit theorem** applies (draw a sufficient number of samples and you end up with the normal distribution)
- These distributions are usually known also as "bell curves" (**Attention:** other distributions take this shape too)

Univariate Standard Normal/Gaussian Distribution

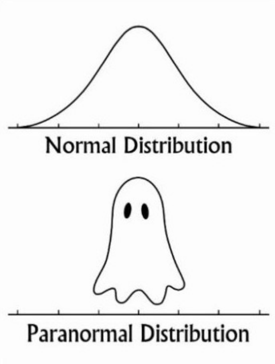
One of the **most important** distributions in natural sciences.

- Used to represent real-valued random variables whose distributions are not known
- The **central limit theorem** applies (draw a sufficient number of samples and you end up with the normal distribution)
- These distributions are usually known also as "bell curves" (**Attention:** other distributions take this shape too)

Univariate Standard Normal/Gaussian Distribution

One of the **most important** distributions in natural sciences.

- Used to represent real-valued random variables whose distributions are not known
- The **central limit theorem** applies (draw a sufficient number of samples and you end up with the normal distribution)
- These distributions are usually known also as "bell curves" (**Attention:** other distributions take this shape too)



Testing For Normality

Testing for normality of the data is **crucial** for certain statistical procedures.

Testing For Normality

Testing for normality of the data is **crucial** for certain statistical procedures.

The Shapiro-Wilks Test In Theory

The QQ Plot In Theory

Testing For Normality

Testing for normality of the data is **crucial** for certain statistical procedures.

The Shapiro-Wilks Test In Theory

- Base assumption: The data is normally distributed
- If p-value $<$ chosen significance level, the data is **not** normally distributed
- Very sensitive to sample size

The QQ Plot In Theory

Testing For Normality

Testing for normality of the data is **crucial** for certain statistical procedures.

The Shapiro-Wilks Test In Theory

- Base assumption: The data is normally distributed
- If p-value $<$ chosen significance level, the data is **not** normally distributed
- Very sensitive to sample size

The QQ Plot In Theory

- Method for comparing two probability distributions by plotting their quantiles against each other
- If the two distributions being compared are similar, the plot will show the line $y = x$.
- Compare the data distribution to the normal distribution

The Shapiro-Wilks Test In R

Using the `shapiro.test()` function:

```r
shapiro.test(rnorm(5000, 20, 2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rnorm(5000, 20, 2)
## W = 1, p-value = 0.7
```

→ Clearly a normal distributed set of values

```r
shapiro.test(seq(1, 500, 5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  seq(1, 500, 5)
## W = 0.95, p-value = 0.002
```

→ Clearly no normal distributed set of values

For data sets bigger than 5000 data points, use the Kolmogorov-Smirnov test (`ks.test()`) in R.

The Shapiro-Wilks Test In R

Using the `shapiro.test()` function:

```
shapiro.test(rnorm(5000, 20, 2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rnorm(5000, 20, 2)
## W = 1, p-value = 0.7
```

→ Clearly a normal distributed set of values

```
shapiro.test(seq(1, 500, 5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  seq(1, 500, 5)
## W = 0.95, p-value = 0.002
```

→ Clearly no normal distributed set of values

For data sets bigger than 5000 data points, use the Kolmogorov-Smirnov test (`ks.test()`) in R.

The Shapiro-Wilks Test In R

Using the `shapiro.test()` function:

```
shapiro.test(rnorm(5000, 20, 2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rnorm(5000, 20, 2)
## W = 1, p-value = 0.7
```

→ Clearly a normal distributed set of values

```
shapiro.test(seq(1, 500, 5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  seq(1, 500, 5)
## W = 0.95, p-value = 0.002
```

→ Clearly no normal distributed set of values

For data sets bigger than 5000 data points, use the Kolmogorov-Smirnov test (`ks.test()`) in R.

The Shapiro-Wilks Test In R

Using the `shapiro.test()` function:

```r
shapiro.test(rnorm(5000, 20, 2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rnorm(5000, 20, 2)
## W = 1, p-value = 0.7
```

→ Clearly a normal distributed set of values

```r
shapiro.test(seq(1, 500, 5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  seq(1, 500, 5)
## W = 0.95, p-value = 0.002
```

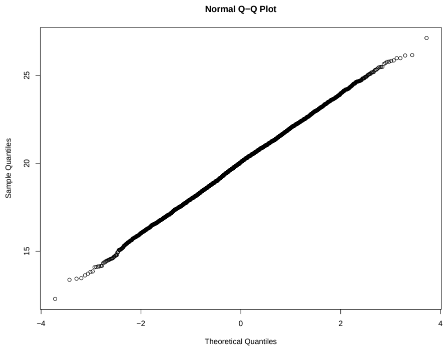
→ Clearly no normal distributed set of values

For data sets bigger than 5000 data points, use the Kolmogorov-Smirnov test (`ks.test()`) in `R`.

The Q-Q Plot

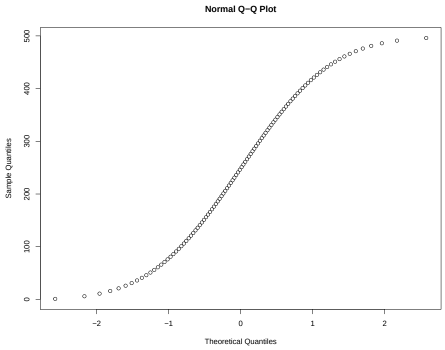
Using the `qqnorm()` function:

`qqnorm(rnorm(5000,20,2))`



→ Clearly a normal distributed set of values

`qqnorm(seq(1,500,5))`

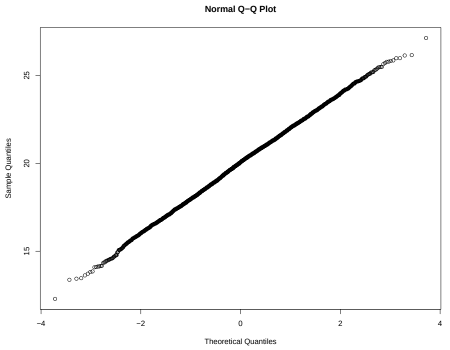


→ Clearly no normal distributed set of values

The Q-Q Plot

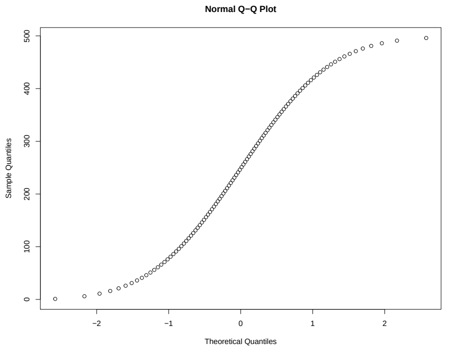
Using the `qqnorm()` function:

```
qqnorm(rnorm(5000,20,2))
```



→ Clearly a normal distributed set of values

```
qqnorm(seq(1,500,5))
```

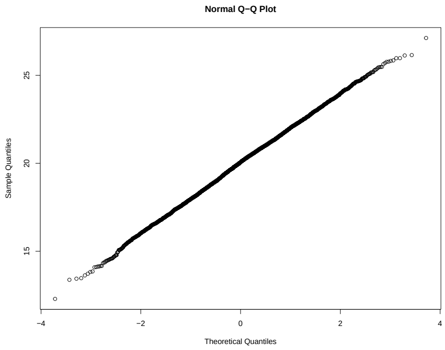


→ Clearly no normal distributed set of values

The Q-Q Plot

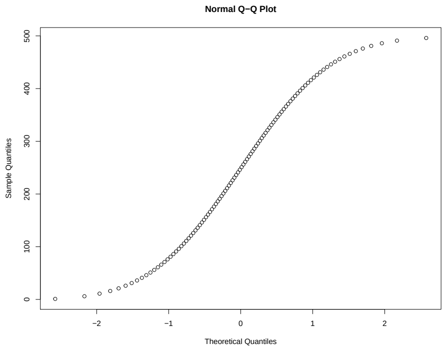
Using the `qqnorm()` function:

```
qqnorm(rnorm(5000,20,2))
```



→ Clearly a normal distributed set of values

```
qqnorm(seq(1,500,5))
```



→ Clearly no normal distributed set of values

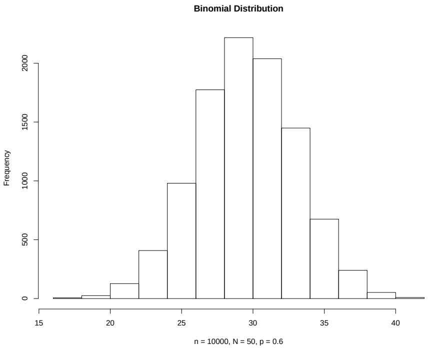
Binomial Distribution

One of the **more important** distributions. It is applicable to:

- Variables which can only take two possible values (e.g. “states”)
- All records of the variable have the same probability *p* of being in one of the two states

It is made up of three **criteria**:

- p - the "success" probability
- n - sample size (how often we sample)
- N - the "binomial total" (for how many individuals we sample each time)



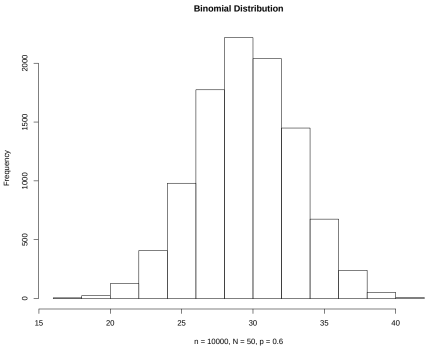
Binomial Distribution

One of the **more important** distributions. It is applicable to:

- Variables which can only take two possible values (e.g. “states”)
- All records of the variable have the same probability *p* of being in one of the two states

It is made up of three **criteria**:

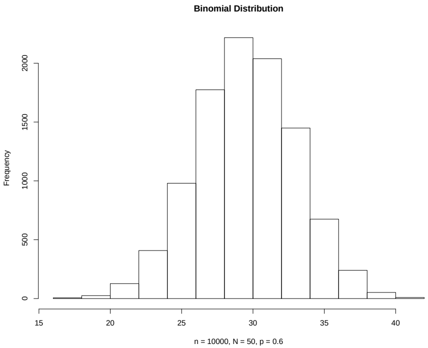
- p - the "success" probability
- n - sample size (how often we sample)
- N - the "binomial total" (for how many individuals we sample each time)



Binomial Distribution

One of the **more important** distributions. It is applicable to:

- Variables which can only take two possible values (e.g. “states”)
- All records of the variable have the same probability *p* of being in one of the two states



It is made up of three **criteria**:

- p - the "success" probability
- n - sample size (how often we sample)
- N - the "binomial total" (for how many individuals we sample each time)

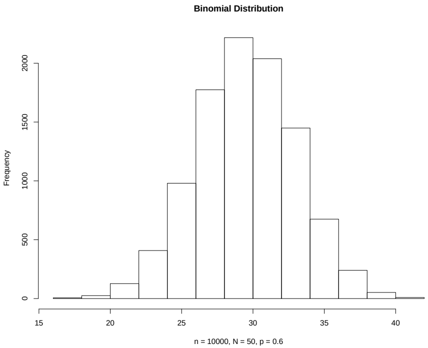
Binomial Distribution

One of the **more important** distributions. It is applicable to:

- Variables which can only take two possible values (e.g. “states”)
- All records of the variable have the same probability *p* of being in one of the two states

It is made up of three **criteria**:

- p - the "success" probability
- n - sample size (how often we sample)
- N - the "binomial total" (for how many individuals we sample each time)



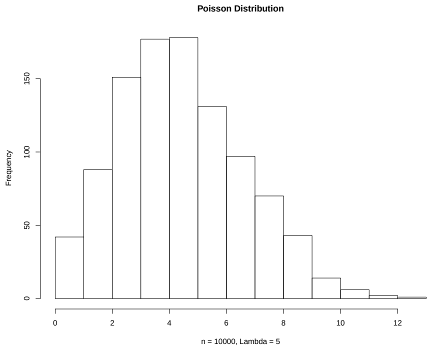
Poisson Distribution

Another one of the **more important** distributions. It is applicable to:

- Focal objects are placed randomly in one or more dimensions
- A random “counting window” (usually one considering time) is placed above the sampling scheme

It is made up of two **criteria**:

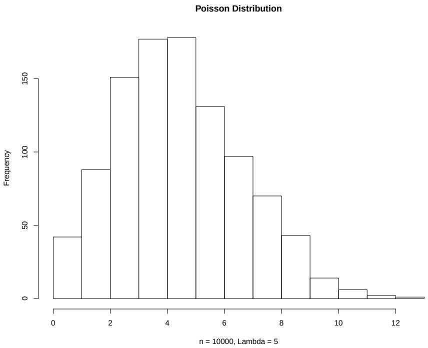
- λ - the mean (= expectation, average count, intensity) as well as the variance (i.e., variance = mean)
- n - sample size



Poisson Distribution

Another one of the **more important** distributions. It is applicable to:

- Focal objects are placed randomly in one or more dimensions
- A random "counting window" (usually one considering time) is placed above the sampling scheme



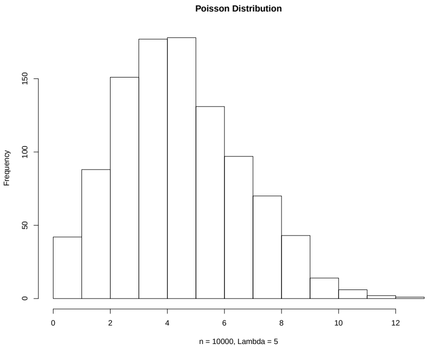
It is made up of two **criteria**:

- λ - the mean (= expectation, average count, intensity) as well as the variance (i.e., variance = mean)
- n - sample size

Poisson Distribution

Another one of the **more important** distributions. It is applicable to:

- Focal objects are placed randomly in one or more dimensions
- A random “counting window” (usually one considering time) is placed above the sampling scheme



It is made up of two **criteria**:

- λ - the mean (= expectation, average count, intensity) as well as the variance (i.e., variance = mean)
- n - sample size

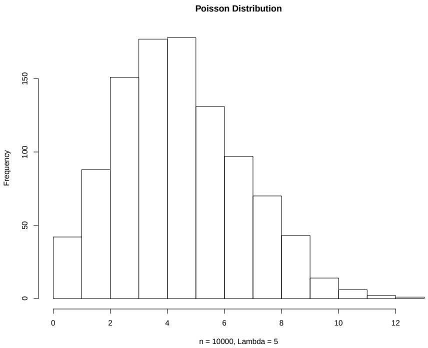
Poisson Distribution

Another one of the **more important** distributions. It is applicable to:

- Focal objects are placed randomly in one or more dimensions
- A random "counting window" (usually one considering time) is placed above the sampling scheme

It is made up of two **criteria**:

- λ - the mean (= expectation, average count, intensity) as well as the variance (i.e., variance = mean)
- n - sample size



How to Measure Distributions

Not all distributions are created equally.

Distributions can be described via **classic parameters of descriptive statistics:**

- Arithmetic Mean
- Mode
- Median
- Minimum, Maximum, Range
- ...

- Variance
- Standard Deviation
- Quantile Range
- **Skewness**
- **Kurtosis**
- ...

How to Measure Distributions

Not all distributions are created equally.

Distributions can be described via **classic parameters of descriptive statistics:**

- Arithmetic Mean
- Mode
- Median
- Minimum, Maximum, Range
- ...

- Variance
- Standard Deviation
- Quantile Range
- **Skewness**
- **Kurtosis**
- ...

How to Measure Distributions

Not all distributions are created equally.

Distributions can be described via **classic parameters of descriptive statistics:**

- Arithmetic Mean
- Mode
- Median
- Minimum, Maximum, Range
- ...

- Variance
- Standard Deviation
- Quantile Range
- **Skewness**
- **Kurtosis**
- ...

How to Measure Distributions

Not all distributions are created equally.

Distributions can be described via **classic parameters of descriptive statistics:**

- Arithmetic Mean
- Mode
- Median
- Minimum, Maximum, Range
- ...

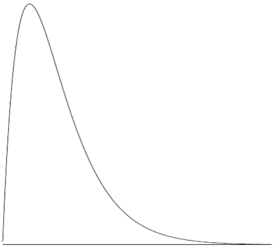
- Variance
- Standard Deviation
- Quantile Range
- **Skewness**
- **Kurtosis**
- ...

Skewness I

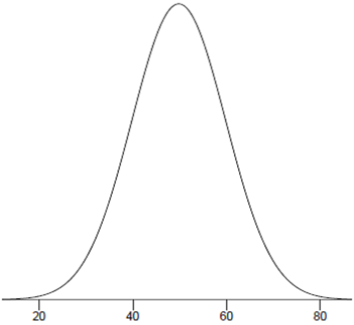
Definition:	Describes the symmetry and relative tail length of distributions.

Positive skew:	Right-hand tail is longer than the left-hand tail
Skew = 0:	Symmetric distribution
Negative skew:	Left-hand tail is longer than the right-hand tail

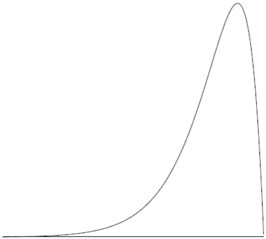
Skewness II



Positive Skew



Symmetric Distribution



Negative Skew

Kurtosis I

Definition:	Describes the evenness/"tailedness" of distributions.
Positive kurtosis:	Short-tailed distribution aka. *leptokurtic*
Kurtosis = 0:	Base representation of a given distribution aka. *mesokurtic*
Negative kurtosis:	Long-tailed distribution aka. *platykurtic*

Kurtosis II