A PRIMER FOR STATISTICAL TESTS

Erik Kusch

erik.kusch@i-solution.de

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

Variables And Their Subsets

What is a variable?

A *variable* presents more or less valuable information about a potential multitude of characteristics of a study system.

What's the fuss?

Sampling effort is limited. We only ever sample a subset of globally present values of any given variable.

So?

Every variable is subject to a certain distribution of its values and each measurement is expected to follow the global distribution to a certain degree.

Variables are, more or less, **raw data**.

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 (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 (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*!

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*.

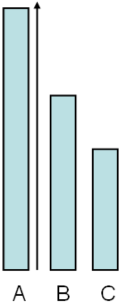
Nominal And Binary

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.

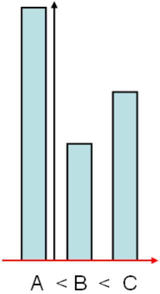
Ordinal

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 mathematical interpretation

Examples:

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



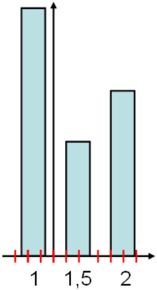
Interval/Discrete

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



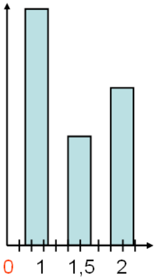
Relation/Ratio

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

Checklist For Your Variables

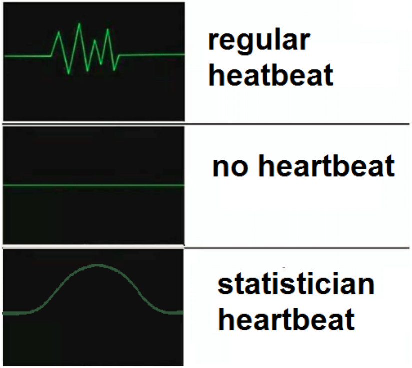
- **What variables** am I using and:
  - Of what *mode* are they?
  - Is every record of the same variable based on the same *unit of measure*?
  - Should I *convert* some of them?

- **What scales** apply and:
  - What do they *imply*?
  - Does my data *fit* these scales?

→ You should be able to **answer these question *before* you begin data collection!**

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.

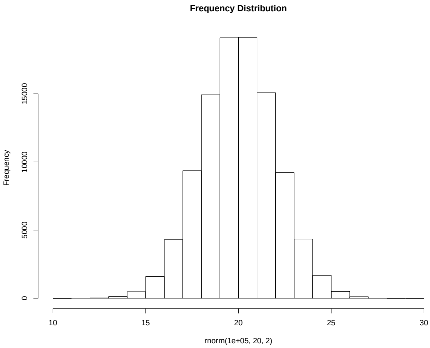


→ Basically **data patterns** we are considering/looking for.

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

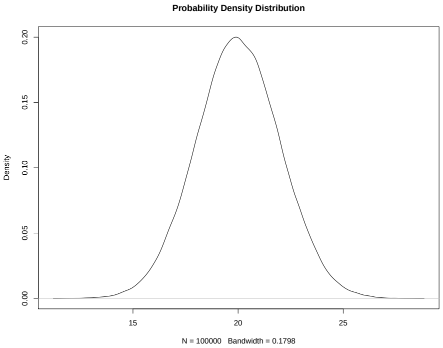



```
hist(rnorm(100000,20,2),
main = "Frequency Distribution")
```

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:

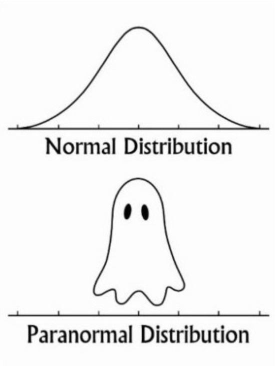
- Area under the curve (AUC) sums to 1
- A probability for every given single value is 0
- The AUC between two values on the X-axis equals the probability to randomly sample a value between these two points

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

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.

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 = 1, p-value = 0.002
```

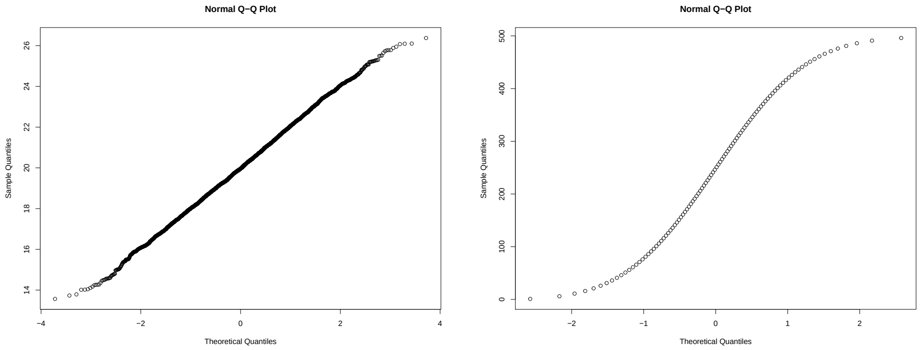
→ Clearly no normal distributed set of values

The Q-Q Plot

Using the `qqnorm()` function:

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

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



→ Clearly a normal distributed set of values

→ Clearly no normal distributed set of values

An Overview of Distributions

There is a **plethora of distributions** which variables could fall onto:

- Bernoulli (probabilities of value 1 and 0 are interdependent)
- **Binomial** (number of successes in a series)
- **Poisson** (probability of a given number of events occurring in a fixed interval of time or space)
- Beta (family of two-parameter distributions with one mode)
- Kent (three-dimensional sphere distribution)
- **Univariate Standard Normal/Gaussian**
- Multivariate Normal
- Log-Normal

. . . and yet **the hat goes deeper**

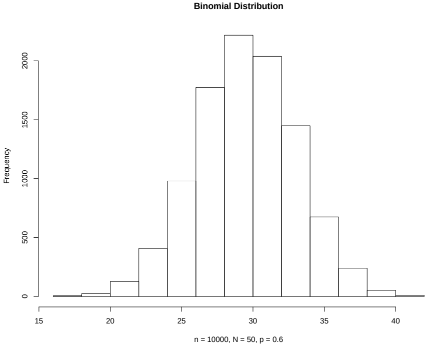
Binomial Distribution

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)



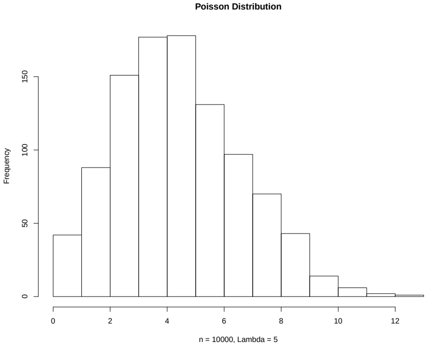
Poisson Distribution

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



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**
- ...

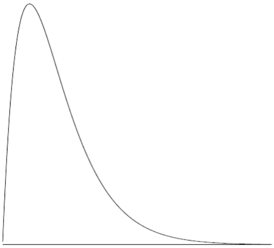
$\rightarrow$ Most of these are dealt with in the next seminar.

Skewness I

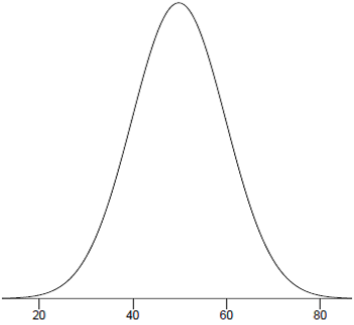
Definition:	Describes the symmetry and relative tail length of distributions.

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

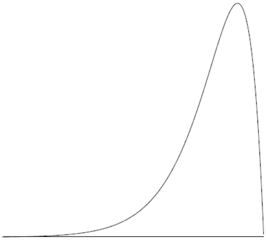
Skewness II



Positive Skew



Symmetric Distribution



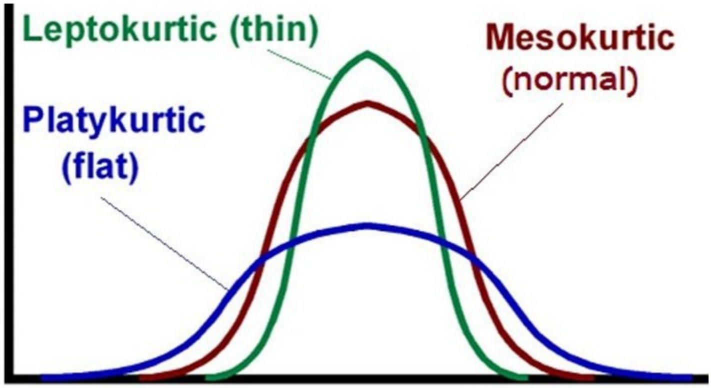
Negative Skew

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



Point and Range Estimations

Point and Range estimations are parameters obtained from a *sample data set* and meant to represent the *population data set*.

- With a probability of 95.5%, the following is true

$$\overline{x} - 2\sigma_{\overline{x}} \leq \mu \leq \overline{x} + 2\sigma_{\overline{x}}$$

- With a probability of 68%, the following is true

$$\overline{x} - \sigma_{\overline{x}} \leq \mu \leq \overline{x} + \sigma_{\overline{x}}$$

$\overline{x}$	Arithmetic mean of the sample
μ	Arithmetic mean of the population
$\sigma_{\overline{x}}$	Standard error of $\overline{x}$

Confidence Intervals I

For multiple confidence intervals (CIs) on a level of α (i.e. 95%), a proportion of α CIs for a given population will contain the arithmetic mean of the population.

$$[\overline{x} - t(\alpha, df) \text{ ; } \overline{x} + t(\alpha, df)]$$

$\overline{x}$	Arithmetic mean of the sample
$\sigma_{\overline{x}}$	Standard error of $\overline{x}$
α	Confidence level (usually 95%)
df	Degrees of freedom
$t(\alpha, df)$	t-value given α and df

Confidence Intervals II

The **Basics of Confidence Intervals**:

- CIs get **larger** when:
  - Smaller sample sizes
  - Bigger spread of data values
  - Higher statistical certainty (α)
- CIs get **narrower** when:
  - Bigger sample sizes
  - Smaller spread of data values
  - Lower statistical certainty (α)

R Environment Objects

R environment objects (stored as `.RData`) are highly valuable objects to any R user because:

- They let you save your entire working environment
- You cannot alter them outside of R (aside from deleting them)

How to **create** them?

- Use the function `save.image()` (you can specify the argument `file` for a specific name of the file)

How to **load** them?

- Use the function `load(...)` ("..." specifies the exact path to the file on your machine)

What to do?

- Load the `Primer.RData` file into R

Variables

Answer the following questions (take a notes for each variable!) for the variables contained within `Primer.RData`:

- **What variables** am I using and:
  - Of what *mode* are they?
- **What scales** apply and:
  - What do they *imply*?
  - Does my data *fit* these scales? Use the functions barplot() and table() when applicable!

Distributions

Plot the distributions of the values for the following variables:

- `Length`
- `Reproducing`
- `IndividualsPassingBy`
- `Depth`

What distributions are these? Use QQPlots or the Shapiro Test to assess normality. If you stumble upon a non-normal distribution, what else could it be?