# ADVANCED BIOSTATISTICS
# ABSTAT18

## Quick Review of Statistical Concepts

Carina Silva
(*carina.silva@estesl.ipl.pt*)
Higher School of Technologies and Health of Lisbon &
Center of Statistics and Applications, University of Lisbon | CEAUL

ESCOLA SUPERIOR DE TECNOLOGIA DA SAÚDE DE LISBOA
INSTITUTO POLITÉCNICO DE LISBOA

CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa

IGC, April 3rd - 6th, 2018

# Introduction

Traditional scientific research consists on four interrelated stages:

- Problem definition.

# Introduction

Traditional scientific research consists on four interrelated stages:

- Problem definition.
- Data gathering.

# Introduction

Traditional scientific research consists on four interrelated stages:

- Problem definition.
- Data gathering.
- Data analysis.

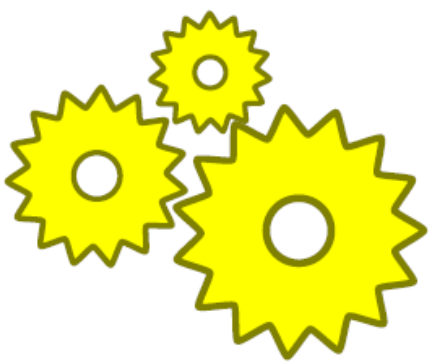
# Introduction

Traditional scientific research consists on four interrelated stages:

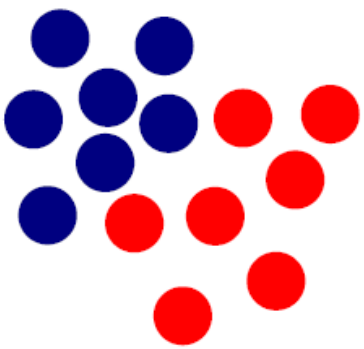
- Problem definition.
- Data gathering.
- Data analysis.
- Data interpretation and conclusions.

# Introduction

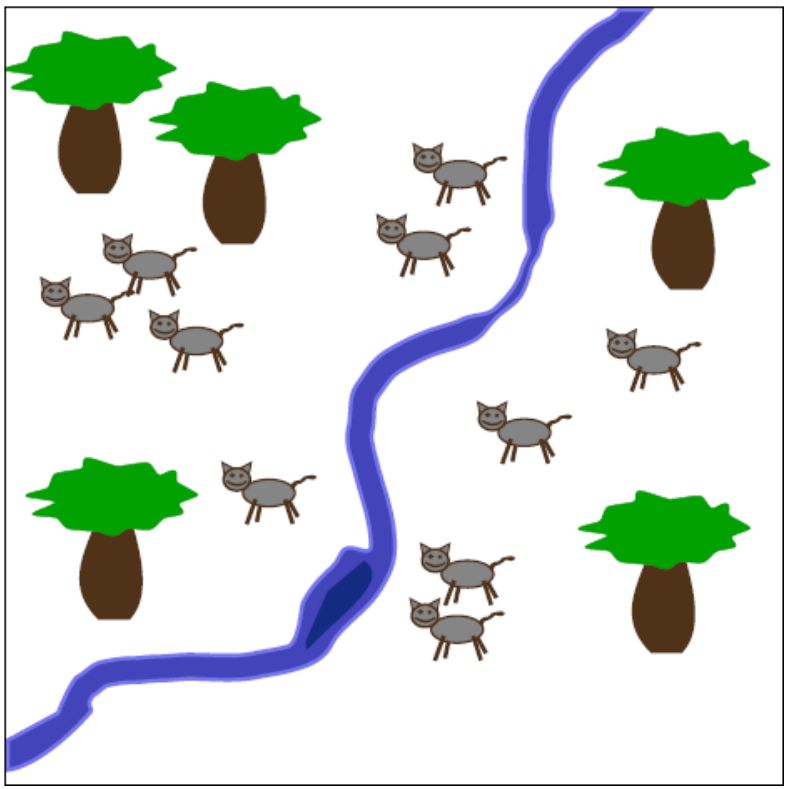
Biological processes



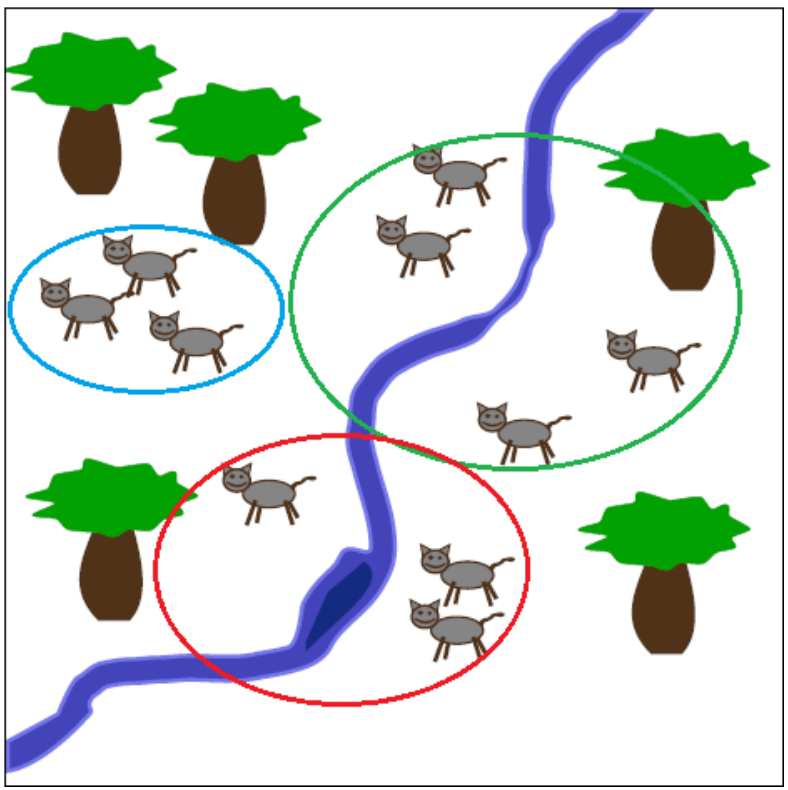
Biological patterns



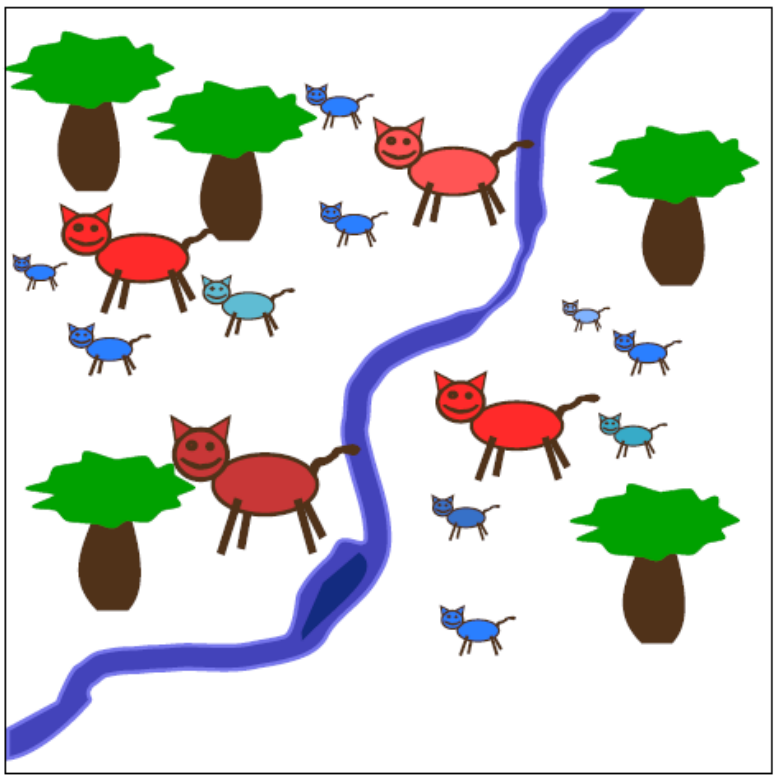
# Identifying genetic groups



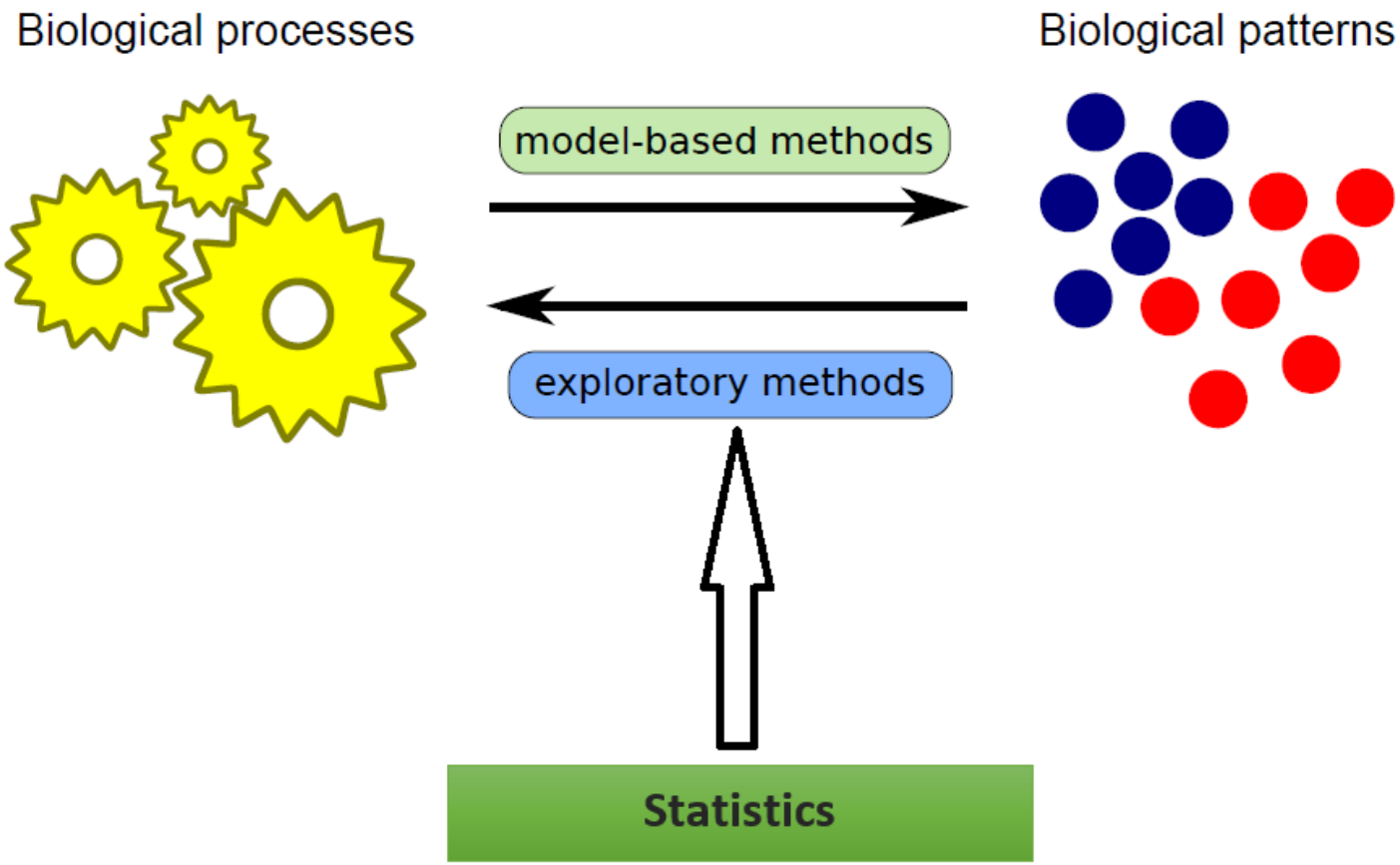
# Identifying genetic groups: Genetic clusters can indicate populations



Correlating genotype and phenotype: Correlations between alleles and phenotypic features can indicate the genetic determinism of a trait.



# Statistics on the rescue!



# Statistics on the rescue!

**Deterministic models:** The equations can be used to determine the value of a specific variable in the model based on the knowledge of the values assumed by other model variables.

$$Y = a + X$$

But there is always some uncertainty in experimental science.

**Probability** used to describe situations were uncertainty occurs.

# Statistics on the rescue!

Thus we need:

**Statistical models:** Allow us to assess the degree of uncertainty present in our experimental results.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

# Probability

There are two principal philosophies to construct statistical models:

| Methodology | Probability viewpoint | Source of Information |
|---|---|---|
| Frequentist (Classical) | Frequentist | Sample data |
| Bayesian | Subjective | Prior and sample data |

# Probability - Kolmogorov's Axioms of Probability

$\Omega$ = sample space, set of all possible outcomes.
$\mathcal{A}$ = set of events
$P$ = The assignment of probabilities to the events

- **Definition of probability**:
  Is a set function $P : \mathcal{A} \to [0, 1]$ which satisfies the axioms
  1. $P[A] \geq 0$ for every $A \in \mathcal{A}$.
  2. $P[\Omega] = 1$.
  3. $A_1, A_2, ...$ sequence of mutually exclusive events in $\mathcal{A}$, then

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} P[A_i] = P[A_1] + P[A_2] + ...$$

- **Definition of Probability space**
  Is the triplet $(\Omega, \mathcal{A}, P[.])$

# Conditional Probability

Let $(\Omega, \mathcal{A}, P[.])$ be a probability space

**Definition: Conditional Probability**

Let $A$ and $B$ two events in $\mathcal{A}$, $P[B] > 0$. Conditional probability of $A$ given $B$

$$P[A|B] = \frac{P[A \cap B]}{P[B]}.$$

# Independence

**Definition: Independent events**

$A$ and $B$ two events in $\mathcal{A}$ are independent if one of the following is satisfied

(i) $P[A \cap B] = P[A]P[B]$

(ii) $P[A|B] = P[A]$ if $P[B] > 0$

(iii) $P[B|A] = P[B]$ if $P[A] > 0$

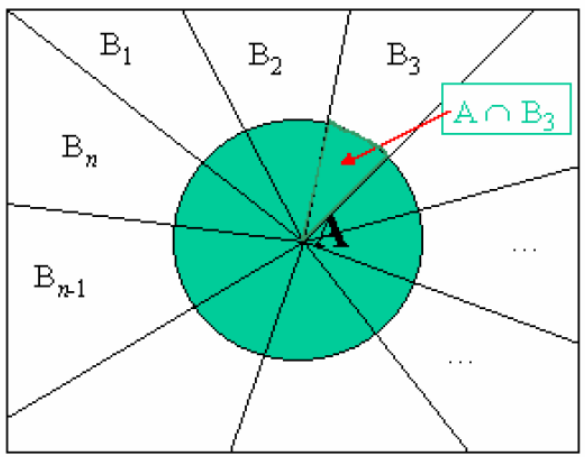
Remark: Independent events can only be mutually exclusive if the probability of at least one is zero.

# Theorem - Theorem of total probabilities

Let $B_1, ..., B_n$ mutually disjoint (or mutually exclusive) in $\mathcal{A}$, $\Omega = \bigcup_{i=1}^{n} B_i$ and $P[B_i] > 0, \forall i$

For every $A \in \mathcal{A}$

$$P[A] = \sum_{i=1}^{n} P[A|B_i]P[B_i].$$



## Theorem - Bayes' Theorem

For every $A \in \mathcal{A}$ such that $P[A] > 0$

$$P[B_k|A] = \frac{P[A|B_k]P[B_k]}{\sum_{i=1}^{n} P[A|B_i]P[B_i]}.$$

Both theorems remain true if $n = \infty$.

# Basic Concepts

- **Variable**: is a characteristic or condition that changes or has different values for different individuals.

# Basic Concepts

- **Variable**: is a characteristic or condition that changes or has different values for different individuals.
- **Observation** (or case): Is a realization of a variable. For example, the weight of a randomly chosen rat is such an observation. (Observations are represented by lowercase, e.g. $x_1$, $x_2$.)

# Basic Concepts

The following definitions are vital in understanding descriptive statistics:

## Types of variables

**Qualitative or Categorical**

- Nominal (race, blood type, etc.)
- Ordinal (the order or rank of the categories is meaningful. For example, staff members may be asked to indicate their satisfaction with a training course on an ordinal scale ranging from "poor" to "excellent".)

**Quantitative**

- Discrete (number of planets orbiting a distant star. Could even be countably infinite.)
- Continuous (your age, not rounded off; weigth)

# Random Variable

- When an experiment is conducted, many characteristics are often measured during and following the experiment.

# Random Variable

- When an experiment is conducted, many characteristics are often measured during and following the experiment.
- The exact numerical value obtained for any particular measurement will depend on many factors such as environmental factors, experimental conditions, or just unexplained factors that always result from taking measurements in real situations.

# Random Variable

- When an experiment is conducted, many characteristics are often measured during and following the experiment.
- The exact numerical value obtained for any particular measurement will depend on many factors such as environmental factors, experimental conditions, or just unexplained factors that always result from taking measurements in real situations.
- We will call such a measurement a **random variable**.

# Random Variable

There are two key parts to the definition of random variable (r.v.):

- The value measured for a random variable must be numeric.
- The value of the random variable that results from an experiment is not known in advance. All we know is that it will be some value from a domain of possible values. Some values are more likely to result than others.

# Random Variable

Example:

Consider the RNA sequence:

AUGCUUCGAAUGCUGUAUGAUGUC

*Using Random Variable X to Quantitatively Model Residues in a Particular RNA Sequence*

| Residue | Value of X (=x) | P (X=x) |
|---|---|---|
| A | 0 | 5/24=0.208 |
| C | 1 | 4/24=0.167 |
| G | 2 | 6/24=0.25 |
| U | 3 | 9/24=0.375 |

# Random Variable

There are two types of random variables:

**Discrete Random Variable** is one that takes a finite distinct values (countable number of possibilities).
Example: $X$ - Number of matching bases when comparing two strands of DNA each of length N. $\mathcal{X} = \{0, 1, ..., N\}$

# Random Variable

**Continuous Random Variable** is one that takes an infinite number of possible values (can take any value on the real line or some subset of the real line).

Example: $X$ - Time until a protein degrades in the cell:,
$\mathcal{X} = (0, \infty)$.

We distinguish discrete from continuous random variables according to whether the sample space $\mathcal{X}$ is countable or not countable.

# Probability Mass Function (pmf)

For a discrete random variable (r.v.) the probability distribution of $X$ is completely determined by specifying $P[A]$ for $A = \{x\}$, for every $x \in \mathcal{X}$.
We represent $P[X = x]$ by $f(x)$. Remember that $\mathcal{X}$ is countable.
It is a function from $\Re$ to $[0, 1]$ such that

1. $f(x) = \begin{cases} \geq 0, & x \in \mathcal{X}; \\ 0, & \text{otherwise.} \end{cases}$
2. $\sum_{x \in \mathcal{X}} f(x) = 1$

To distinguish the different values $x$ in $\mathcal{X}$ we can write $x_i$.

Example: Suppose that an individual is the progeny from crossing randomly two lines, and that each line consists of genotypes $A_1$ and $A_2$. Let $X$ be the r. v. representing the number of $A_2$ alleles in the progeny from this cross. From Mendel's laws, the probability distribution of $X$ is:

| Genotype | $A_2A_2$ | $A_2A_1$ | $A_1A_1$ |
|---|---|---|---|
| $x$ | 2 | 1 | 0 |
| $p(x)$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |

# Cumulative distribution function

The **cumulative distribution function** (cdf), or just **distribution function**, describes the probability that a real-valued random variable $X$ with a given probability distribution will be found at a value less than or equal to $x$. Intuitively, it is the "area so far" function of the probability distribution.

**Discrete cdf**

$$F_X(x_k) = P(X \leq x_k) = \sum_{i=1}^{k} f(x_i),\ k \leq n$$

Example: Suppose that an individual is the progeny from crossing randomly two lines, and that each line consists of genotypes $A_1$ and $A_2$. Let $X$ be the r. v. representing the number of $A_2$ alleles in the progeny from this cross. From Mendel's laws, the probability distribution of $X$ is:

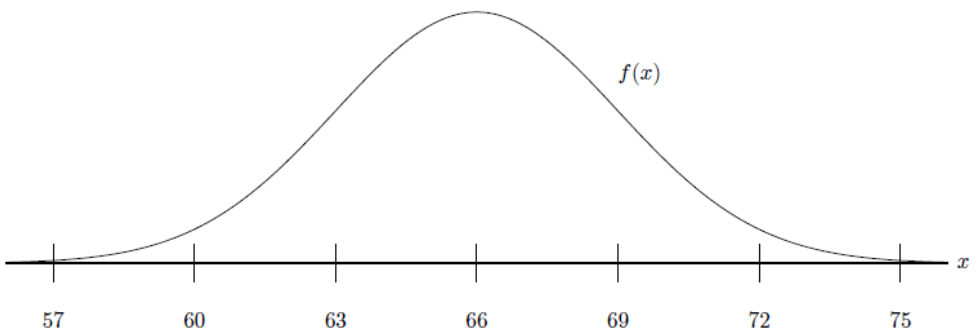
| Genotype | $A_2A_2$ | $A_2A_1$ | $A_1A_1$ |
|---|---|---|---|
| $x$ | 2 | 1 | 0 |
| $p(x)$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |

The probability distribution function is

$$F(x) = \begin{cases} 0, & \text{for } x < 0; \\ \frac{1}{4}, & \text{for } 0 \leq x < 1; \\ \frac{3}{4}, & \text{for } 1 \leq x < 2; \\ 1, & \text{for } x \geq 2. \end{cases}$$

# Probability Density Function - pdf

- A probability density function (pdf), is a function associated with a **continuous random variable**
- Areas under pdf correspond to probabilities for that random variable



# Probability Density Function - pdf

- To be a valid pdf, a function $f$ must satisfy
  1. $f(x) \geq 0$ for all $x$
  2. $\int_{-\infty}^{+\infty} f(x)dx = 1$

# Cumulative Distribution Function - cdf

- **Definition**: *Cumulative distribution function (cdf)*
  $F_X : \Re \to [0, 1]$, satisfying

$$F_X(x) = P[X \leq x]$$

  for every real $x$.

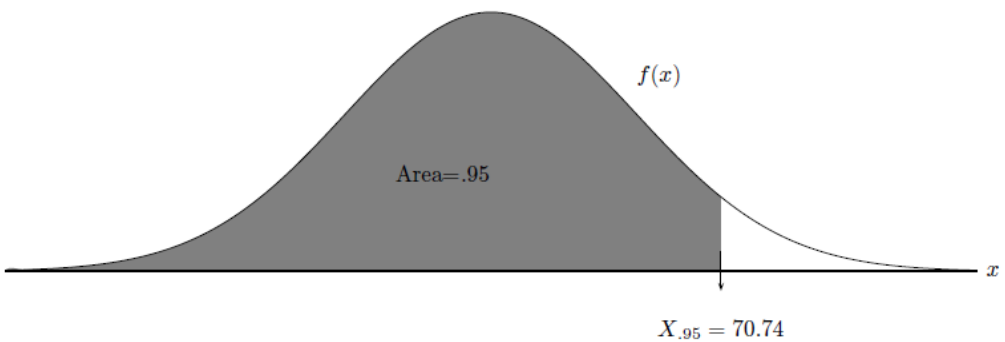
$F_X$ describes the probability that a random variable $X$ with a given probability distribution will be found at a value less than or equal to $x$.

# Quantiles

The $\alpha^{th}$ **quantile** of a distribution with cumulative distribution $F$ is the point $x_\alpha$:
$F(x_\alpha) = \alpha$
A **percentile** is simply a quantile with $\alpha$ expressed as a percent.
The **median** is the $50^{th}$ percentile.

# Quantiles

Example: Suppose that 70.74 inches is the 95th percentile of height according to a probability model. It means that 95% of all adults have a height less than or equal to 70.74 inches, and so 5% are taller than 70.74 inches.
In notation, this says that $\chi_{.95} = F_X^{-1}(.95) = 70.74$ or $P(X \leq 74.70) = .95$



# Parameters

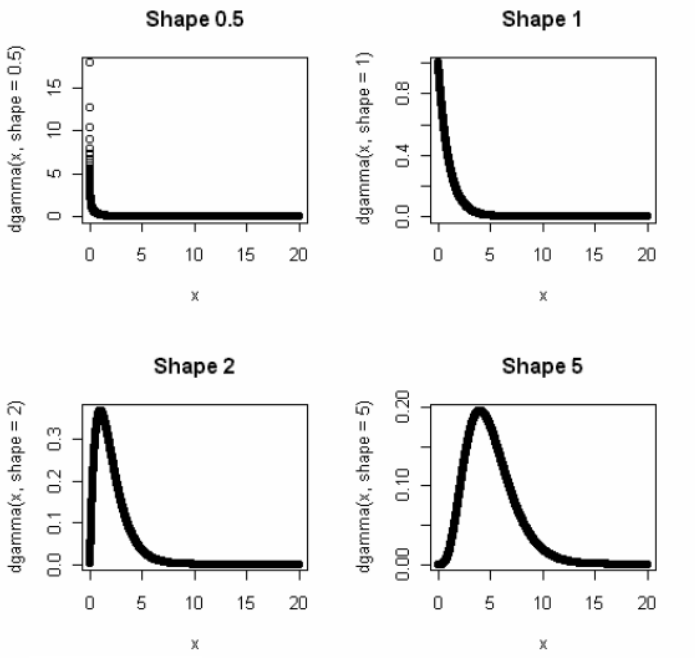
- The most general definition of a **parameter** is some constant involved in a probability distribution.
- Random variables define the data in a probability model.
- Parameters serve to mathematically frame how a probability model fits.

# Parameters

- Parameters represent characteristics of the model they are used in and usually are classified as different types, such as shape, scale and location.
- The idea is you want a distribution to fit your data model "just right" without a fit that is "overfit".

# Shape parameter

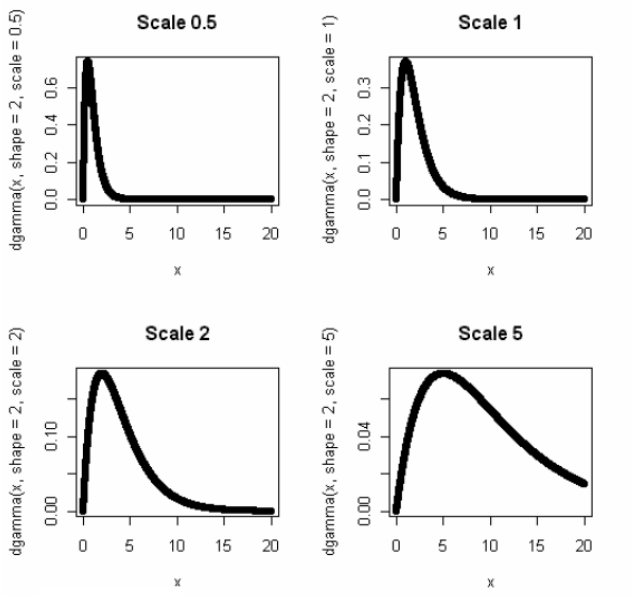
- The same data is used but modeled using different values for the shape parameter for this distribution (with all other parameters constant). Notice that changing the shape parameter changes how the model fits the data.




*Altering a Shape Parameter with Other Parameters Held Constant*

# Scale parameter

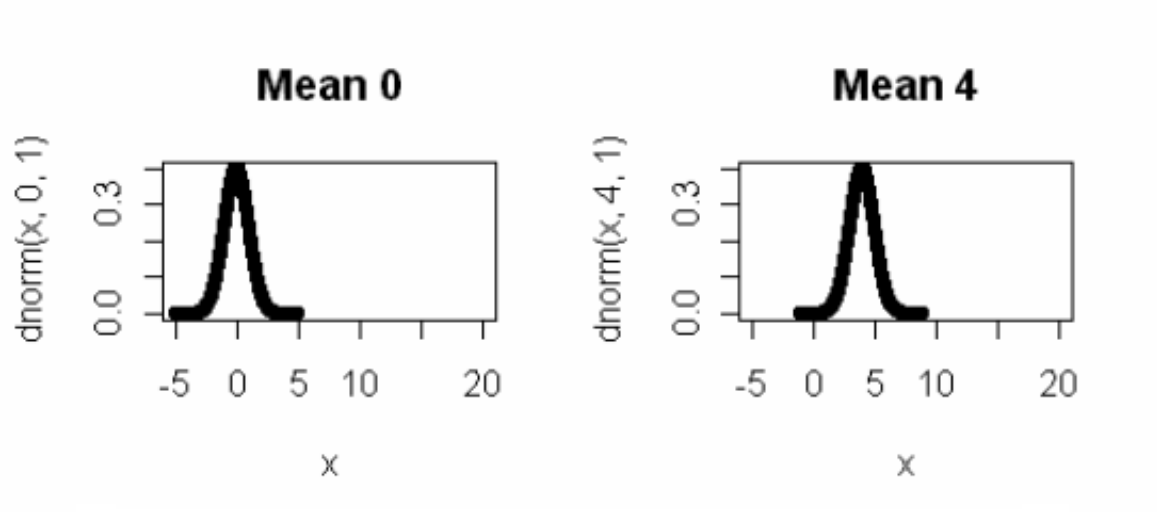
- Scale parameters do not change the shape of a distribution but change how spread out the distribution is.




*Altering a Scale Parameter with Other Parameters Held Constant*

# Location parameter

- The location parameter specifies where the graph lies along the horizontal (x) axis.




*The Effect of changing the Location Parameter*

# Exercises

Task: Do exercises 1 to 5 of the `probreview.md` file.

# Parameter Estimation

There are mainly two problems to be solved:

- Given a probability model, validate its correctness based on how well it describes real data.
- Given a probability model that we assume to be correct, estimate the unknown parameters of the model based on data that has been collected (or conduct tests regarding those parameters).

## Given a probability model, validate its correctness based on how well it describes real data

- Let $X$ be a continuous random variable that measures the amount of time it takes for a certain cellular protein to degrade.
- $X$ is certainly a random variable since any particular observation of this process will yield different times due to random and unknown factors in the cellular process.
- We are interested in constructing a probability model that does a good job of describing degradation for this particular protein.

# Given a probability model, validate its correctness based on how well it describes real data

- Let's start out considering the possible models.
- Since $X$ is a time measurement, we are likely to consider probability models that take this into account (i.e., allow for the random variable to take on non-negative real numbers).
- Very common distributions to use to model waiting times (in this case, waiting time until the protein degrades), can be gamma distribution and is special case the exponential distribution.

# Given a probability model, validate its correctness based on how well it describes real data

- For this example, let's say that we will use the exponential distribution. So, we formally define our model like this:

$$X_1, X_2, ..., X_n \sim Exp(\lambda)$$

- This states that we will take $n$ independent observations of this random variable, and that measured values will follow an exponential distribution with unknown parameter $\lambda$.
- So, our first goal is to estimate $\lambda$ based on data we collect.

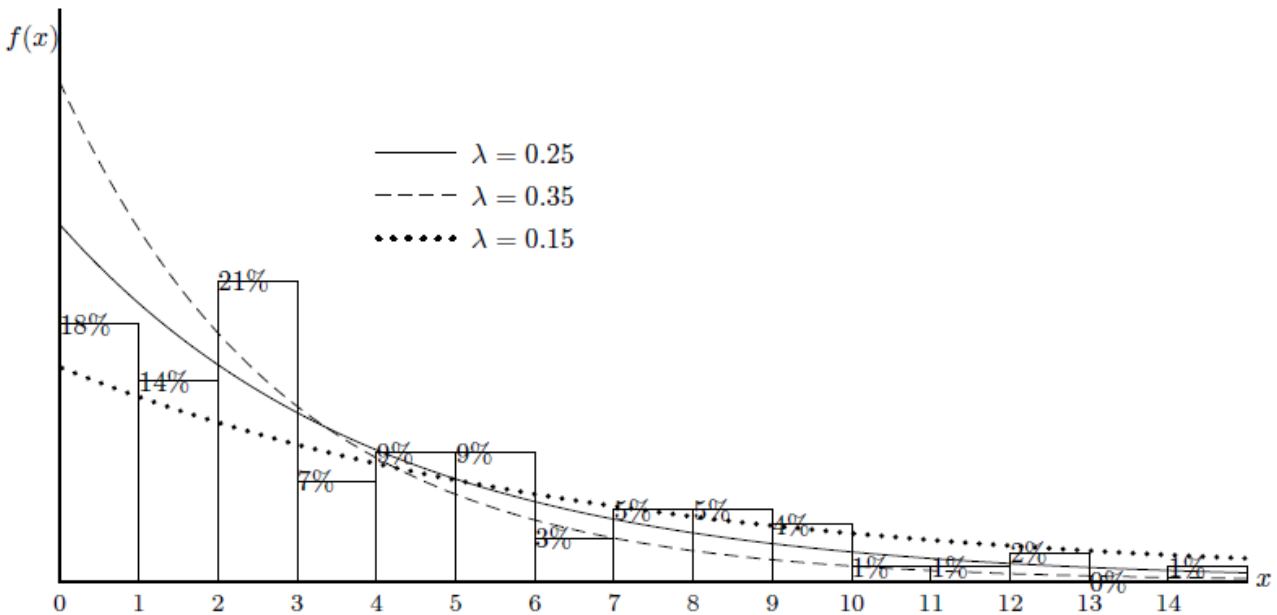
# Given a probability model, validate its correctness based on how well it describes real data

- Say we collect data on $n = 100$ randomly selected such proteins.

| | | | | | | |
|---|---|---|---|---|---|---|
| 1.0488 | 8.7128 | 5.1088 | 2.4621 | 8.0735 | 4.4189 | 4.2664 |
| 5.0919 | 5.5437 | 1.6770 | 1.7151 | 1.7572 | 0.5448 | 2.6872 |
| 4.2220 | 1.0151 | 5.3031 | 3.8807 | 2.0359 | 0.0426 | 1.4084 |
| 6.7384 | 0.1361 | 2.0712 | 3.3359 | 0.6761 | 2.7327 | 4.2882 |
| 4.4953 | 1.8219 | 1.2427 | 8.9060 | 12.2685 | 2.8841 | 0.5807 |
| 0.5927 | 3.4842 | 0.9928 | 9.4022 | 3.8196 | 6.3884 | 2.1940 |
| 0.2338 | 2.1911 | 0.9576 | 3.9605 | 2.7559 | 6.6619 | 2.2530 |
| 4.2422 | 0.7012 | 1.1716 | 5.0964 | 1.8256 | 0.4665 | 0.9726 |
| 2.0984 | 7.5901 | 1.1320 | 7.7666 | 7.2670 | 9.8356 | 1.1392 |
| 0.5069 | 11.2457 | 5.7494 | 10.2586 | 2.5182 | 1.9329 | 9.8432 |
| 2.0279 | 0.6773 | 8.9189 | 2.3682 | 8.1927 | 7.2889 | 7.9581 |
| 12.2379 | 2.9023 | 2.3436 | 2.4534 | 0.7263 | 1.4295 | 0.8745 |
| 2.9707 | 0.4734 | 3.8200 | 3.3817 | 5.2861 | 5.1474 | 0.4276 |
| 5.7718 | 4.3280 | 2.0507 | 2.2834 | 14.4425 | 4.1508 | 9.9144 |
| 2.1074 | 4.4184 | | | | | |

# Given a probability model, validate its correctness based on how well it describes real data

- What is the best estimate for the parameter $\lambda$?



## Given a probability model that we assume to be correct, estimate the unknown parameters of the model based on data that has been collected

We don't just eyeball it when truly solving this problem. But this is the essence of the problem of parameter estimation - attempting to determine the value of unknown parameters that seems to fit best with the data we collected. In reality, we will rely on techniques like the **method of moments (MOM)** or **maximum likelihood estimation (MLE)** to solve the problem.

# Maximum Likelihood Estimation

## Algorithm

- (1) Write down the joint pdf of the random sample. We will also refer to this as the likelihood function of the parameter(s), and view it as a function of the parameter(s) instead of a function of the random variables. Call this function $L(\theta)$ where $\theta$ is our generic symbol for the parameter of the model.
  - Consider the exponential distribution example, where we have $X_1, X_2, ..., X_n \sim Exp(\lambda)$, the likelihood function will be: $L(\lambda) = f_{(X_1,X_2,...X_n)}(x_1, x_2, ..., x_n) = f_{x_1}(x_1) \times ... \times f_{X_n}(x_n) = (\lambda e^{-\lambda x_1})...(\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \sum x_i}$

# Maximum Likelihood Estimation

- (2) Take the natural *log* of $L$, giving us $lnL(\theta)$.
  - $lnL(\lambda) = nln - \lambda \sum x_i$

# Maximum Likelihood Estimation

- (3) Differentiate with respect to each parameter individually (or just once if there is only one parameter).
  - $\frac{dlnL(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum x_i$

# Maximum Likelihood Estimation

- (4) Set the resulting equation(s) equal to zero and solve for the parameters. The result gives you the MLE for these parameters. (This last step can be impossible in some situations.)
  - $\frac{n}{\lambda} - \sum x_i = 0 \Longleftrightarrow \hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{3.938} = 0.2539$

# Example

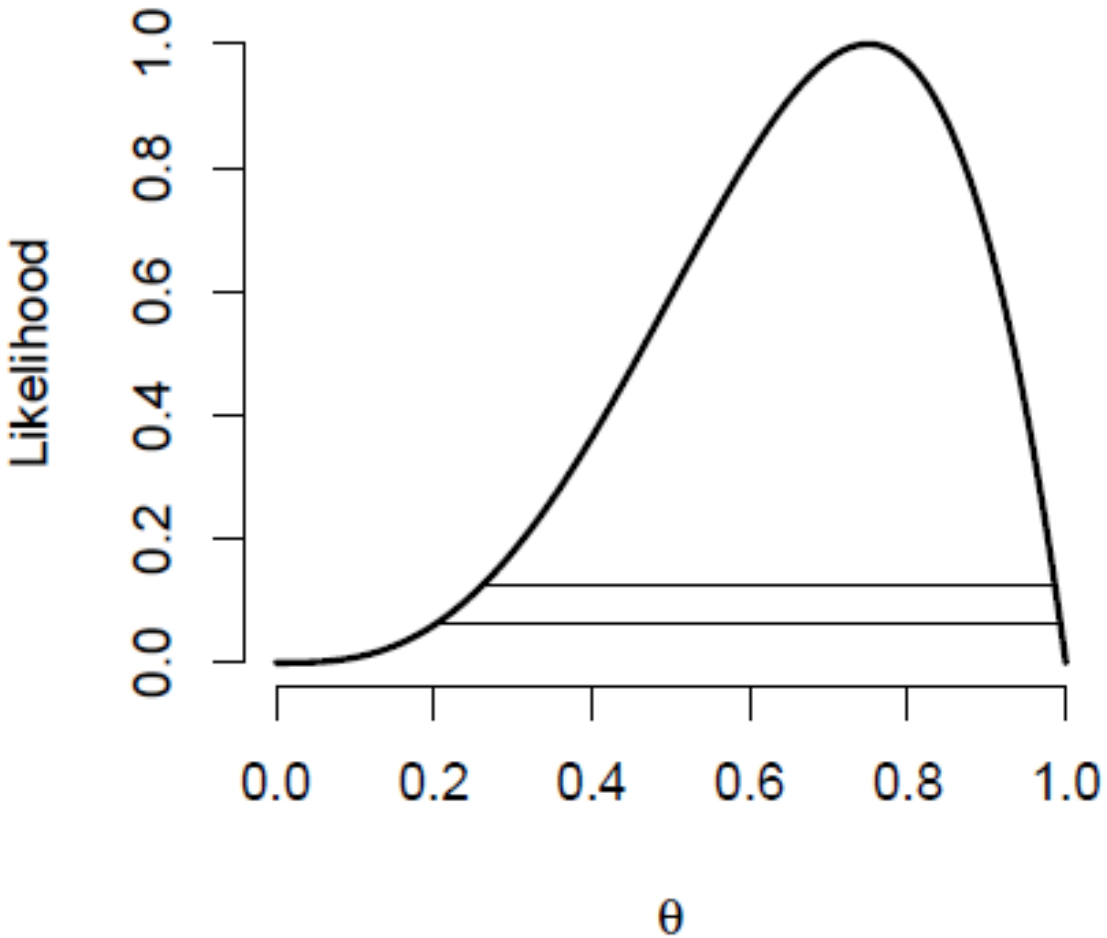
- Suppose that we flip a coin with success probability $\theta$
- Recall that the mass function for $X$ is $f(x, \theta) = \theta^x(1-\theta)^{1-x}$ for $\theta \in [0, 1]$
  where $x$ is either 0 (Tails) or 1 (Heads)
- Suppose that the result is a Head
- The likelihood is:
  $\mathcal{L}(\theta, 1) = \theta^1(1-\theta)^{1-1} = \theta$ for $\theta \in [0, 1]$
- Therefore, $\mathcal{L}(.5, 1)/\mathcal{L}(.25, 1) = 2$
- There is twice as much evidence supporting the hypothesis that $\theta = 0.5$ to the hypothesis that $\theta = 0.25$

# Example (cont)

- Suppose now that we flip our coin from the previous example 4 times and get the sequence 1, 0, 1, 1
- The likelihood is:
  $\mathcal{L}(\theta, 1, 0, 1, 1) =$
  $\theta^1(1-\theta)^{1-1}\theta^0(1-\theta)^{1-0}\theta^1(1-\theta)^{1-1}\theta^1(1-\theta)^{1-1} = \theta^3(1-\theta)^1$
  for $\theta \in [0, 1]$

# Plotting likelihoods

- Generally, we want to consider all the values of $\theta$ between 0 and 1
- A likelihood plot displays $\theta$ by $\mathcal{L}(\theta, x)$
- Usually, it is divided by its maximum value so that its height is 1


Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
θ

# Maximum likelihood

- The value of $\theta$ where the curve reaches its maximum has a special meaning
- It is the value of $\theta$ that is most well supported by the data
- This point is called the maximum likelihood estimate (or MLE) of $\theta$:
  $MLE = argmax_\theta \mathcal{L}(\theta, x)$
- Another interpretation of the MLE is that it is the value of $\theta$ that would make the data that we observed most probable.

# Exercises

Task: Exercise 6 of the `probreview.md` file