Introduction to probability and statistics

Vincent Guillemot
Hugo Varet
Some helpful documents

Statistiques pour Statophobes – Denis Poinsot (link)

Points of significance – Nature website (link)

Seeing theory (link)


Why statistics? Because…

• It takes into account the random nature of objects
• In biology, every individual/culture/sample is different!
• It helps making decisions on objective criteria rather than on a subjective perception

... think about girafes
Feed the 3rd reviewer
Different aspects of a statistical study

1. Collecting data
   Sampling, experiment design

Two very different samples can come from the same population.

Two similar samples can come from two different populations.
Different aspects of a statistical study

2. Exploratory analysis
   Data structure, suggest hypotheses
Different aspects of a statistical study

3. Inferential statistics (statistical induction)
Hypothesis testing, estimation

Random sampling

Blue : 1/6 (8/41)
Black : 1/6 (5/41)
Red : 2/6 (6/41)
Green : 2/6 (22/41)
Different aspects of a statistical study

4. Statistical modeling
   Regression, classification (machine learning), ...
Course outline

1. Introduction, random variables and point estimation
2. Sampling, estimation, confidence intervals
3. Hypothesis testing
4. Multivariate data analysis: PCA, clustering
5. Experimental design
6. Project
Quick reminders and notations
Notations

Let a sample of $n$ values:

$$x_1, x_2, x_3, ..., x_n$$

We can write:

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} = \sum_{i=1}^{10} x_i$$
RStudio reminder
Exercise (1): R reminder

1. Use the function `data()` to load the built-in dataset `cars`.
2. Can you see it in the environment?
3. Type the instruction `View(cars)`?
4. Can you see the dataset now in the environment?
5. What does the instruction `cars$dist` do?
6. What does the instruction `cars$dist >= 30` do?
7. What does the instruction `cars[cars$dist >= 30, ]` do?
8. What does the instruction `sum(cars$dist >= 30)` do?
9. What is the minimum distance for a speed > 15?
Probability versus statistics
Formal definitions

PROBABILITY

Probability is the measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1 (where 0 indicates impossibility and 1 indicates certainty).

STATISTICS

Statistics is a set of methods aiming at describing, analyzing and interpreting, in a quantified manner, phenomena that are by nature composed of numerous events that can be counted and classified.
Definitions with hats

**PROBABILITY**

All the **universe** is known

**STATISTICS**

Only a **sample** is known
Why do we need *formal* probability?

A cab is involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

1. 85% of the cabs in the city are Green and 15% are Blue.
2. A witness has identified the cab as Blue.
3. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

**Question:** what is the probability that the cab involved in the accident was Blue rather than Green?

From Tversky and Kahneman (1982)
85% of green cabs

15% of blue cabs
85% of green cabs

15% of blue cabs

Witness said “blue”
Why do we need *formal* probability?

Assume there are 100 cabs:

<table>
<thead>
<tr>
<th>Reality</th>
<th>Green</th>
<th>Blue</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blue</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

What the witness says

- The witness is right XXX times over the YYY "blue" cabs he saw.

\[
\text{Proba(cab is blue | witness said "blue")} = \frac{XXX}{YYY} = ZZZ\%
\]

Typical answer without using formal probabilities: 80%
With equations!

$H$ (Hypothesis): “the cab is blue”

$D$ (data): “the witness says the cab is blue”

$P(D) = P(D \cap H) + P(D \cap \overline{H})$

$= P(D|H)P(H) + P(D|\overline{H})P(\overline{H})$

$= 0.8 \times 0.15 + 0.2 \times 0.85$

$= 0.29$

Now,

$P(H|D) = P(D|H) \frac{P(H)}{P(D)}$

$= 0.8 \times \frac{0.15}{0.29} \approx 0.41$
Trees
Factorial function

\[ n! = 1 \times 2 \times \cdots \times n \]

Is a function that can be generalized to a continuous space (it is then noted \( \Gamma \)):

\[
\Gamma: z \mapsto \int_0^{+\infty} t^{z-1}e^{-t} \, dt ;
\]

We can show that

\[ \forall n \in \mathbb{N}, \Gamma(n + 1) = n! \]
Quick exercise (2)

1. Look for the factorial function in R.
2. What R package does the function you found belong to?
Combinations

Let $E$ be a set of $n$ elements.

The number of combinations of $k$ elements of $E$ is called the binomial coefficient:

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Order doesn’t matter and no repetition is allowed!
Small example

Let $E = \{ A, B, C \}$ a set of 3 elements. We can extract all the combinations of 2 elements:

$$\{A, B\} - \{A, C\} - \{B, C\}$$

**Note**: $\{A, B\}$ is similar to $\{B, A\}$ as order does not matter.

The number of combinations of 2 elements among 3 is:

$$C_3^2 = \frac{3!}{2!(3-2)!} = 3$$
R interlude: build your own function

GENERAL STRUCTURE

\[
f \leftarrow \text{function(} \text{arg1, arg2, ...}\}\{ \\
\text{Command 1} \\
\text{Command 2} \\
... \\
\text{return(result)} \\
\}
\]

EXAMPLE

\[
poly2 \leftarrow \text{function(} x, a=1, b=0, c=0\}\{ \\
\text{res} \leftarrow a \times x^2 + b \times x + c \\
\text{return(res)} \\
\}
\]
Exercise (3)

Create a function that:

1. Takes as arguments two integers \( n \) and \( k \),
2. Computes the number of combinations of \( k \) elements among \( n \):

\[
\binom{n}{k} = \frac{n!}{k!(n-k)!}
\]

There are 7 cars that are in parking violation in the “rue Michel”, but the police representative only has 4 tickets left.

How many different “fining” possibilities are there?

*(Use your own function and then the `choose()` function)*
Remarks on creating functions in R

• Functions are objects
• Functions have local environments
• Good practices:
  • use named arguments
  • use the simplest output possible
  • indent!
  • do not use global variables
Probability
Blaise Pascal (1623 – 1662)
But also...

Gerolamo Cardano
Abraham de Moivre
Pierre Simon Laplace
Marin Mersenne
Pierre de Fermat
Antoine Gombaud. Better known as: “Le Chevalier de Méré”
What was the object under study?
So what is Probability?

Let’s start with the universe, i.e. a set of events:

**Probability**: assign a value between 0 and 1 to any combination of events, e.g.:

“get a red circle” or “get either a red circle or a blue triangle”
Probability: an intuitive definition

\[ P(A) = \frac{\text{Number of cases where } A \text{ is true}}{\text{Number of all possible cases}} \]
P stands for...

- Probability of the sample space: $P(\Omega) = 1$
- Probability of an empty set: $P(\emptyset) = 0$
- Union: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Complement: $P(\overline{A}) = P(A^C) = 1 - P(A)$
- Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Independent events: $P(A \cap B) = P(A)P(B)$
- Incompatibles events: $P(A \cap B) = 0$
Tossing 4 coins

Outcome Probability

HHHH $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

HHHT $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

HHTH $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

HHTT $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

HTHH $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

HTHT $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

HTTH $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

HTTT $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

THHH $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

THTH $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

THTT $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

TTHH $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

TTHT $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

TTTH $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$

TTTT $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$
Exercise (4)

• Create a vector $x$ containing the probabilities of having 0, 1, 2, 3 or 4 heads, i.e. $1/16, 4/16, 6/16, 4/16, 1/16$
• Check that the sum of its elements is equal to 1.
• Use the `barplot()` function on it to obtain the probability distribution.
Interlude: barplots with colors

barplot(height = x, col="steelblue")  barplot(height = x, col=1:4)

Now try this: colors()
Random variable
Non-random variable

Question: 4 apples cost 2€, how much does 1 apple cost?

Equation:

$$4 \times apple\text{Price} = 2$$

$$apple\text{Price} = 2/4$$

$$apple\text{Price} = 0.5€$$

Looking for the apple price considered as fixed.
Non-random variables

**System** of two equations with two unknown variables:

\[
\begin{align*}
3x - y &= 1 \\
2x + 3y &= 19
\end{align*}
\]

Result:

\[x = 2 \quad \text{and} \quad y = 5\]
Random variables

10 values are obtained by measuring a biological phenomenon:

4, 6, 3, 7, 7, 3, 2, 9, 1, 2

Let $X$ be a random variable describing this biological phenomenon:

- $x_1 = 4$ is the 1$^{st}$ realization of $X$
- $x_2 = 6$ is the 2$^{nd}$ realization of $X$
- ...
- $x_{10} = 2$ is the 10$^{th}$ realization of $X$
Random variables

If you redo the experiment:

- You will get 10 different values/realizations...
- ...but the random variable $X$ remains unchanged

We cannot plan/guess exactly the result of the experiment
Statistical distribution

**Goal**: find the most appropriate description for the random variable $X$ from the observed data

- An infinity of statistical laws exists
- Some of the most famous are...
  - Bernoulli, Binomial, Poisson...
  - Uniform, Gaussian, Exponential...
Different types of variables

**Quantitative**: numerical data on which various mathematical operations can be performed (sum, average, ...)
  - **Continuous** data take continuous values within an interval
  - **Discrete** data take discrete values in a finite set of possible values
Different types of variables

Qualitative:

- **Categorical (nominal) data**
  - Binary: male/female, TRUE/FALSE
  - Eye color

- **Ordinal** or ranked data are not numerical but can be ordered
  - Small - medium - large

Quantitative data can be transformed into qualitative data, e.g. age:

- < 20
- Between 20 and 30
- Between 30 and 40
- > 40
Discrete random variable

Probability function that associates to each event a probability:

- A single probability is a positive real number between 0 and 1
- The sum of all the probabilities is equal to 1
Bernoulli

Two possible outcomes: 0 (failure) or 1 (success)

Parameter:

\[ p = \text{probability of 1 (success)} \]

Probability table:

<table>
<thead>
<tr>
<th>Probability</th>
<th>Get 0</th>
<th>Get 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - p</td>
<td></td>
<td>p</td>
</tr>
</tbody>
</table>

- Toss: heads or tails
Quick exercise (5)

Concrete example: Toss a coin!

What is the probability of “success”? 

Smallest R exercise: take this opportunity to draw the corresponding probability function:
Binomial

Possible outcomes: 0, 1, 2, 3, ..., n

Parameters:
- \( p \) = probability of success
- \( n \) = number of experiments

Probability table:

\[
P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.
\]

Concrete example:
- Launch \( n = 10 \) times a coin and count the number \( k \) of heads
Exercise (6)

- Go to the help page of the `rbinom()` function in R.
- Generate a vector containing \( n=100 \) realizations of a Binomial random variable of parameters \( \text{size}=10 \) and \( \text{prob}=0.5 \).
- What is the most observed realization?
A binomial experiment

Keeeweee or Kooowooo?
The result of Chastaing (1958)

**Hypothesis:** without any other prior knowledge, we tend to associate meaning to words based on their phonological properties:

“eee” would mean small and “ooo” big.

**Result:** 9 children out of 10 agreed with Chastaing’s theory.

What is the 3rd reviewer going to say?
Uniform

Possible outcomes: 1, 2, 3, ... or k

Parameters:

- $k = \text{number of possible outcomes}$
- $p = \text{prob. of getting } 1 = \ldots = \text{prob. of getting } k = \frac{1}{k}$

Probability table:

<table>
<thead>
<tr>
<th></th>
<th>Get 1</th>
<th>Get 2</th>
<th>...</th>
<th>Get $k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>$\frac{1}{k}$</td>
<td>$\frac{1}{k}$</td>
<td></td>
<td>$\frac{1}{k}$</td>
</tr>
</tbody>
</table>
Example: dice
Continuous random variable

**Probability density** associating a probability to an interval:
- it’s a positive function,
- its integral is equal to 1.
Uniform

Parameters:
- left endpoint,
- right endpoint.

All the intervals of the same length are equiprobable.

Example: distribution of p-values under the null hypothesis.
Gaussian

Two parameters:
- Mean $\mu$
- Variance $\sigma^2$

Examples:
- Height
- Central Limit Theorem.
Discrete vs continuous

<table>
<thead>
<tr>
<th></th>
<th>Discrete law</th>
<th>Continuous law</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Possible values</strong></td>
<td>1, 2, 3, 4, 5, 6 or all the positive integers or all the integers</td>
<td>All real numbers or real numbers in [a, b]</td>
</tr>
<tr>
<td><strong>What we compute</strong></td>
<td>$\text{Prob}(X = k)$ with $k$ in the list of possible values</td>
<td>$\text{Prob}(x_1 \leq X \leq x_2)$ with $x_1$ and $x_2$ chosen</td>
</tr>
<tr>
<td><strong>Constraint</strong></td>
<td>$\sum_{k=0}^{n} \text{P}(X = k) = 1$</td>
<td>Area under the density curve is equal to 1</td>
</tr>
</tbody>
</table>
Theoretical quantiles

The $q^{th}$ theoretical quantile corresponds to an actual value $Q$ which separates the universe in two sub-groups such that:

• $q\%$ of the observed values are lower than $Q$
• $(1 - q)\%$ are greater than $Q$
In R

You can go from quantile to probability with the function 

\textit{pdistribution(quantile)}

And you can go from probability to quantile with the function 

\textit{qdistribution(probability)}

\textbf{Examples:}
\begin{align*}
punif(q=15, \min=10, \max=20) \\
quunif(p=0.5, \min=10, \max=20)
\end{align*}
What else?

`rdistribution()` allows to generate random numbers from the given distribution.

`ddistribution()` allows to get the density of the given distribution.
Example (7): Gaussian proba’s

Use the `pnorm()` function to compute the probability that a Gaussian variable with mean 1 and standard deviation equal to 2 is smaller than 4.
Example (8): Gaussian quantiles

Use the `qnorm()` function to compute the 5% quantile of a Gaussian distribution with mean 0 and standard deviation 1.

How about the 2.5% quantile?

NB: “5%” means 0.05 and “2.5%” means 0.025
Chi-2 distribution

A chi² variable is defined as the sum of \( k \) squared independent gaussian variables with parameters (0,1).

Only one parameter:
- number of degrees of freedom \( k \).

**Example**: salary!
Quick exercise (9)

1. Generate a random \( \chi^2 \) sample with 5 degrees of freedom of size 100 with the function `rchisq()`.
   Call it \( x \).

2. Generate 5 random normal variables (i.e., Gaussian with mean 0
   and standard deviation 1) of the same size with `rnorm()`.
   Call them \( z_1, z_2, z_3, z_4 \) and \( z_5 \).

3. Create a new sample called \( y \) equal to the sum of the 5 squared
   normal samples.

4. Use `hist()` to compute the empirical distribution of both \( x \) and
   \( y \): how similar are they?