Permutation tests

Fabian Pedregosa

October 3, 2017

Data Science Learn2Launch, UC Berkeley

Announcements

- **Next week is the first presentation!**
  1. 10 min presentation (by teams) + 5 min questions
  2. At least: objective of the project, dataset, exploratory analysis.
- Server, more CPUs, GPUs, etc $\implies$ register at AWSEducate: `https://www.awseducate.com/Registration`. If this is not enough, come and see me.
- Office hours: me 3pm-5pm SDH 421, Bowen Mondays on demand.

Structure of this lecture

- Me: explain the method of permutation tests.
- You: solve problem based on this method.
- You: volunteer presents his solution, gets +0.5 point bonus (out of 10) on final grade.
- Me: Introduction to supervised learning. Logistic regression.

Permutation tests

We will answer the burning question

Does drinking beer make you more attractive to mosquitos?

OPEN ACCESS Freely available online

PLoS one

Beer Consumption [redacted] Human Attractiveness to Malaria Mosquitoes

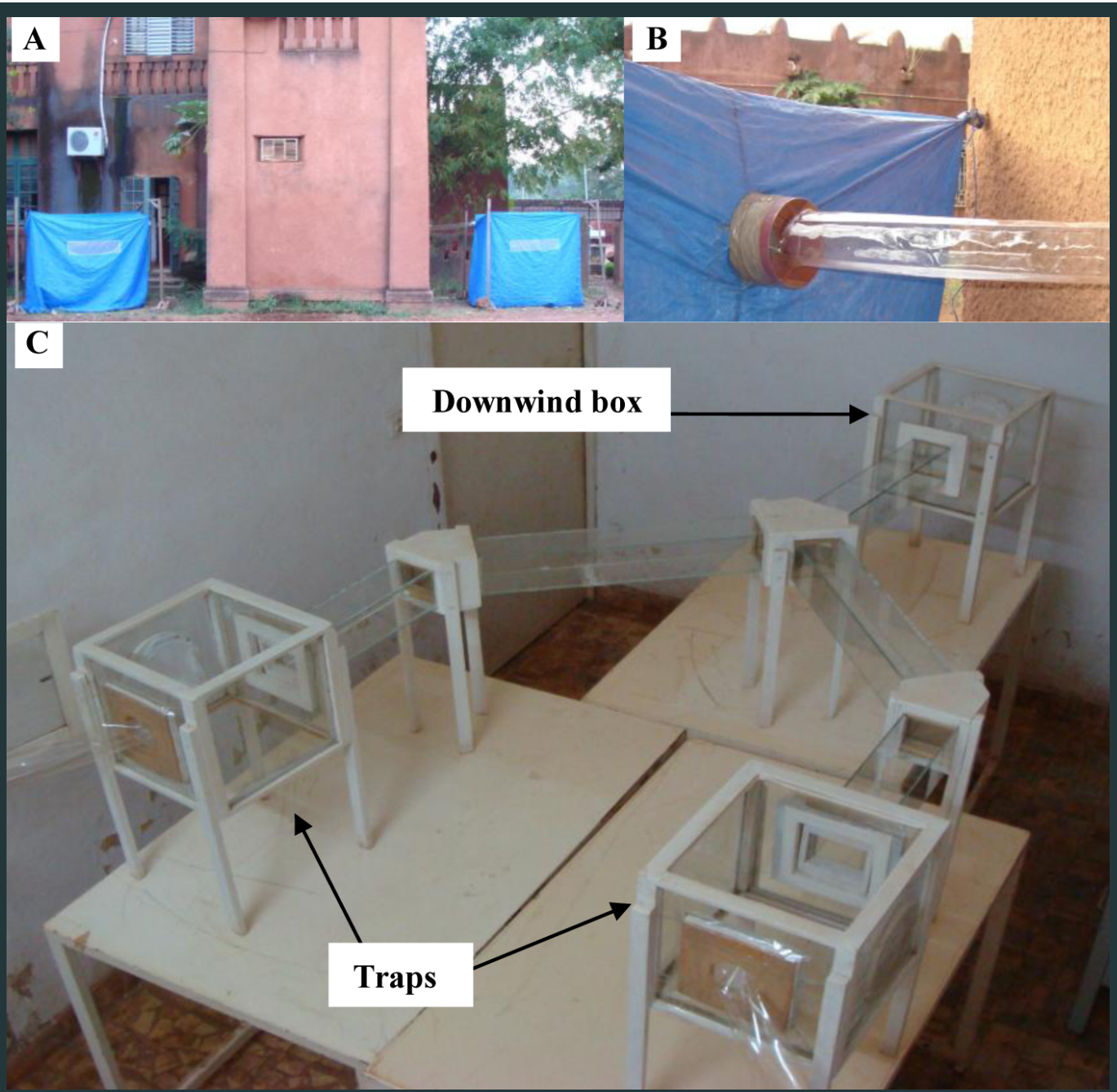
Thierry Lefèvre[1]*, Louis-Clément Gouagna[2,3], Kounbobr Roch Dabiré[3,4], Eric Elguero[1], Didier Fontenille[2], François Renaud[1], Carlo Costantini[2,5], Frédéric Thomas[1,6]

1 Génétique et Evolution des Maladies Infectieuses, UMR CNRS/IRD 2724, Montpellier, France, **2** Caractérisation et Contrôle des Populations de Vecteurs, IRD/UR 016, Montpellier, France, **3** Institut de Recherche en Science de la Santé, Bobo-Dioulasso, Burkina Faso, **4** Laboratoire de Parasitologie et d'Entomologie Médicale, Centre Muraz, Bobo-Dioulasso, Burkina Faso, **5** Organisation de Coordination pour la lutte contre les Endémies en Afrique Centrale, Yaoundé, Cameroun, **6** Institut de Recherche en Biologie Végétale, Université de Montréal, Montréal, Canada

Abstract

Background: Malaria and alcohol consumption both represent major public health problems. Alcohol consumption is rising in developing countries and, as efforts to manage malaria are expanded, understanding the links between malaria and alcohol consumption becomes crucial. Our aim was to ascertain the effect of beer consumption on human attractiveness to malaria mosquitoes in semi field conditions in Burkina Faso.

Experiment



Data

Beer	Water
27 19 20 20 23 17 21 24 31 26 28 20 27 19 25 31 24 28 24 29 21 21 18 27 20	21 19 13 22 15 22 15 22 20 12 24 24 21 19 18 16 23 20
$\text{mean}_{\text{beer}} = 23.6$	$\text{mean}_{\text{water}} = 19.2$

$$\text{mean}_{\text{beer}} - \text{mean}_{\text{water}} = 4.4$$

Statistical problem

Is the difference of 4.4 sufficient to claim that drinking beer makes you more attractive to mosquitos?

What is the probability of this happening by chance? $\Longrightarrow$ Statistical problem.

Null hypothesis (H_0), both means are equal and the difference is due to chance.

Instances of this problem are pervasive in data science: does an upgrade increase user engagement?, is the new algorithm generating more revenue? is the new treatment effective? etc.

Two approaches: *i*) Statistics 101 and *ii*) computational method.

Statistics 101

Stats 101

- t-test

Stats 101

- t-test
- Test statistic: $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}}$, where $s_p = \sqrt{\frac{s^2_{X_1} + s^2_{X_2}}{2}}$

Stats 101

- t-test
- Test statistic: $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}}$, where $s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$
- Which under the null hypothesis follows a Student t distribution

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Stats 101

- t-test
- Test statistic: $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}}$, where $s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$
- Which under the null hypothesis follows a Student t distribution

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- $\nu =$ degrees of freedom

Stats 101

- t-test
- Test statistic: $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}}$, where $s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$
- Which under the null hypothesis follows a Student t distribution

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- $\nu =$ degrees of freedom The degrees of freedom ν is approximated using the Welch–Satterthwaite equation

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\nu_1} + \frac{s_2^4}{N_2^2\nu_2}}$$

Stats 101

- t-test
- Test statistic: $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}}$, where $s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$
- Which under the null hypothesis follows a Student t distribution

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- $\nu =$ degrees of freedom The degrees of freedom ν is approximated using the Welch–Satterthwaite equation

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\nu_1} + \frac{s_2^4}{N_2^2\nu_2}}$$

Stats 101

- t-test
- Test statistic: $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}}$, where $s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$
- Which under the null hypothesis follows a Student t distribution

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Skeptic: I don't believe this!

- $\nu =$ degrees of freedom The degrees of freedom ν is approximated using the Welch–Satterthwaite equation

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\nu_1} + \frac{s_2^4}{N_2^2\nu_2}}$$

Computational method

Data

Beer	Water
27 19 20	21 19 13
20 23 17	22 15 22
21 24 31	15 22 20
26 28 20	12 24 24
27 19 25	21 19 18
31 24 28	16 23 20
24 29 21	
21 18 27	
20	
$\text{mean}_{\text{beer}} = 23.6$	$\text{mean}_{\text{water}} = 19.2$

$$\text{mean}_{\text{beer}} - \text{mean}_{\text{water}} = 4.4$$

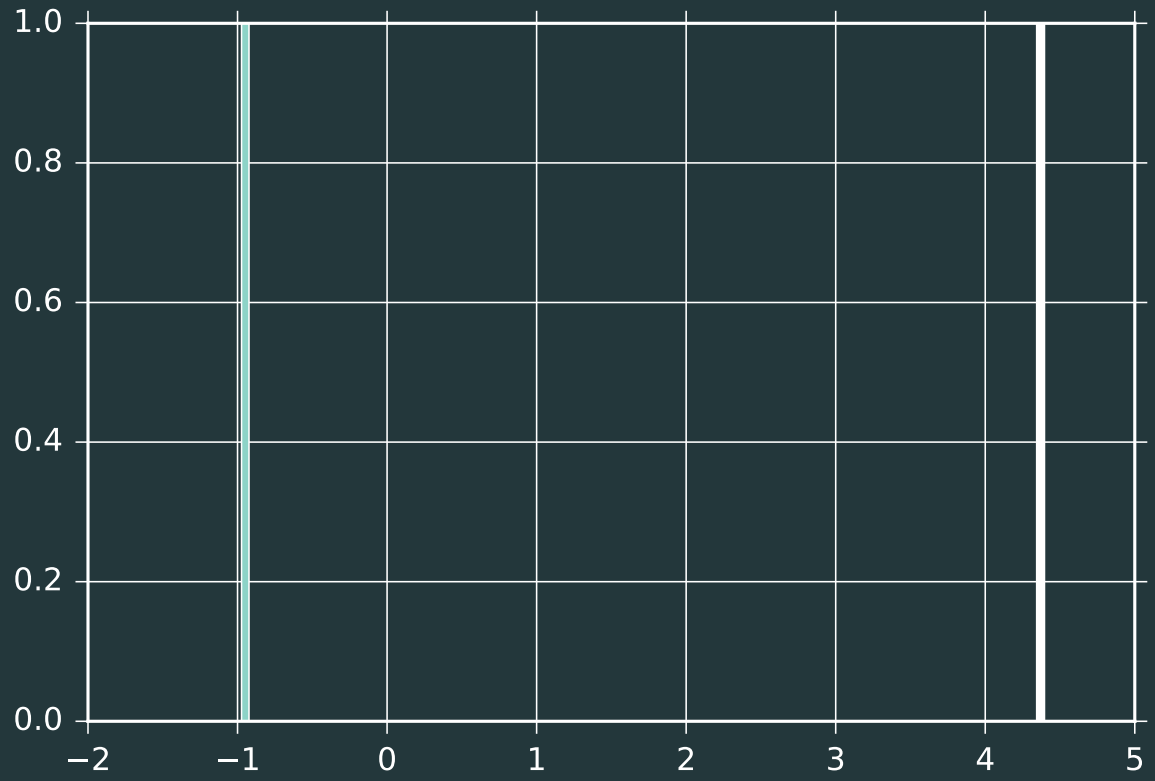
Data

Beer	Water
21 19 20	27 19 27
15 23 17	22 20 22
21 24 31	15 22 20
26 28 20	12 24 24
27 19 25	23 19 27
31 24 28	16 21 20
24 29 21	
17 18 27	
20	
$\text{mean}_{\text{beer}} = X$	$\text{mean}_{\text{water}} = Y$

$$\text{mean}_{\text{beer}} - \text{mean}_{\text{water}} = -0.9$$

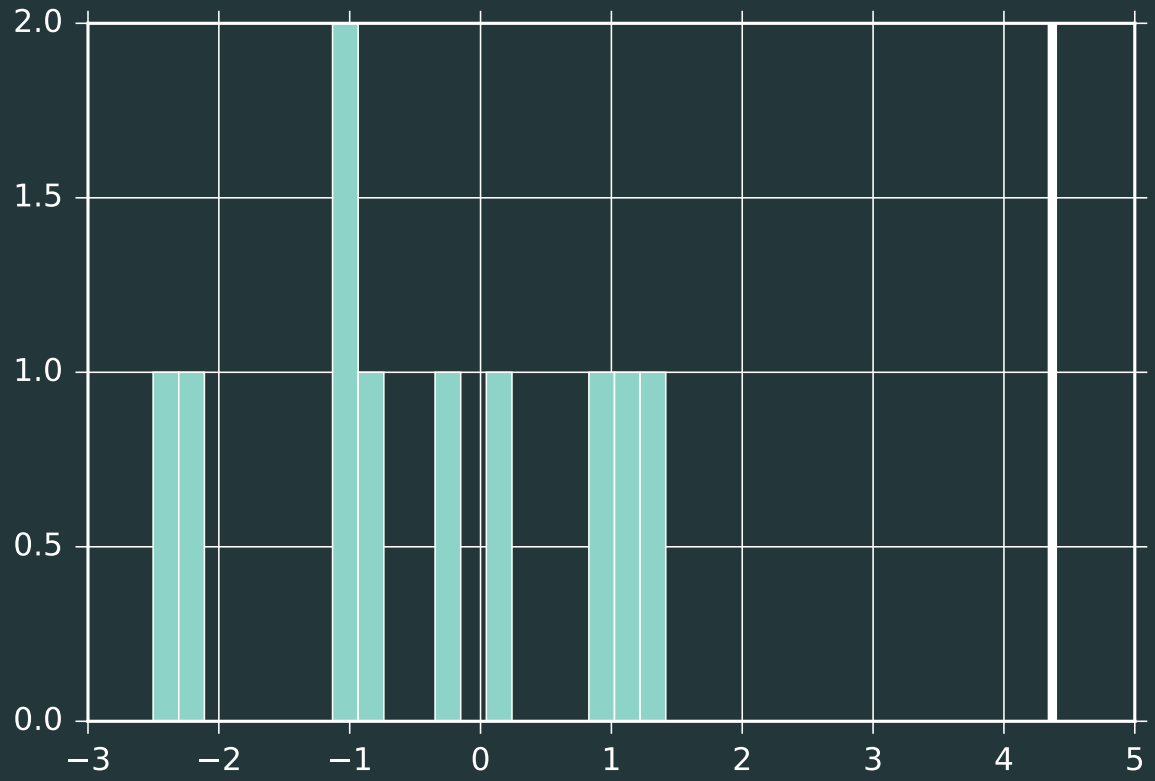
Data

1 permutation



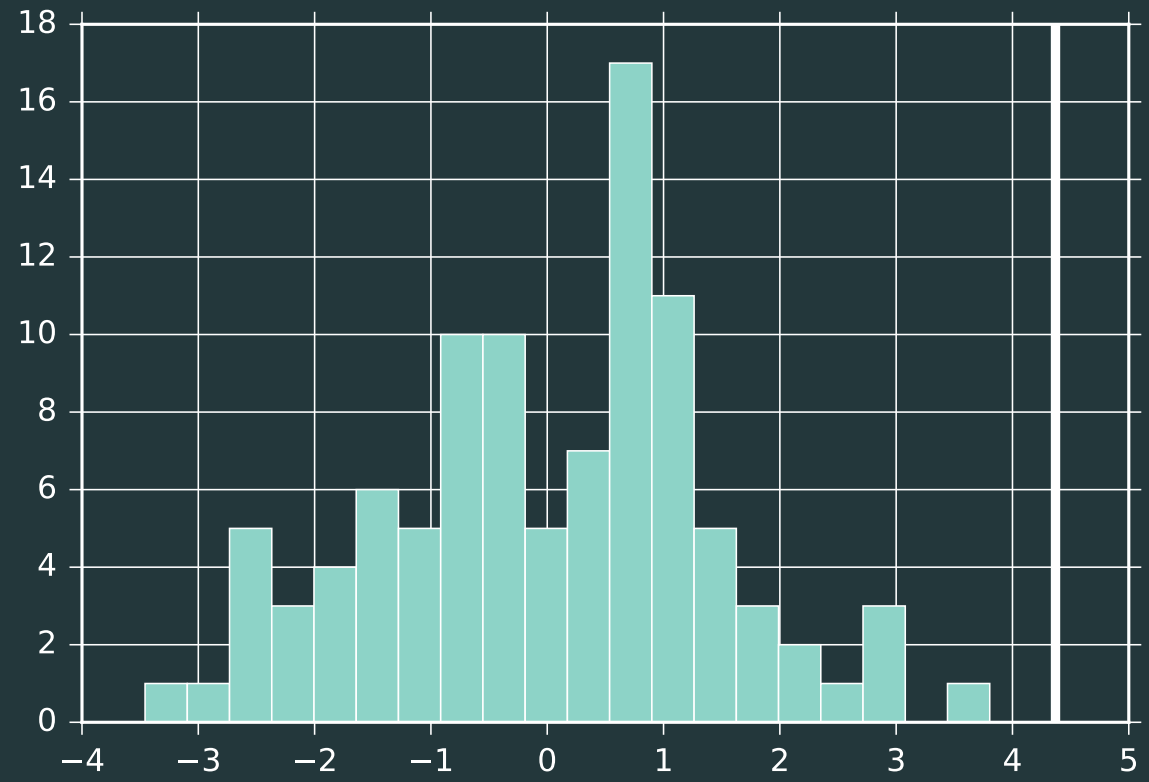
Data

10 permutation



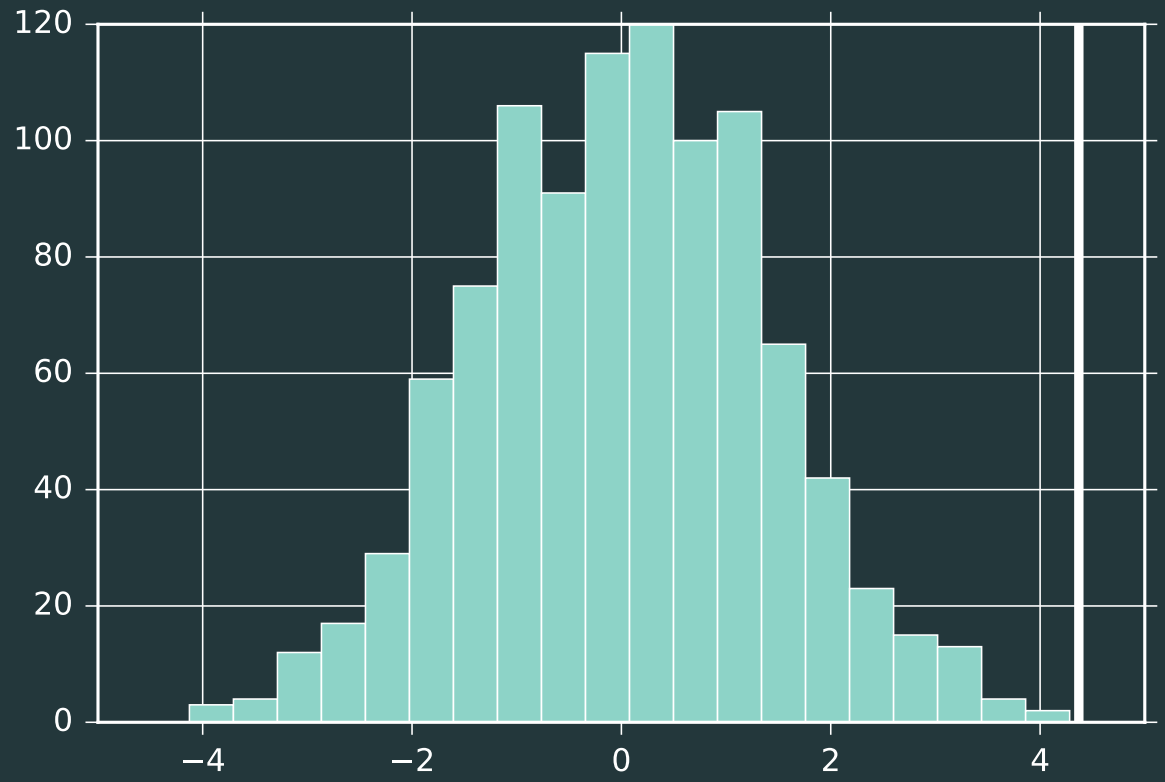
Data

100 permutation



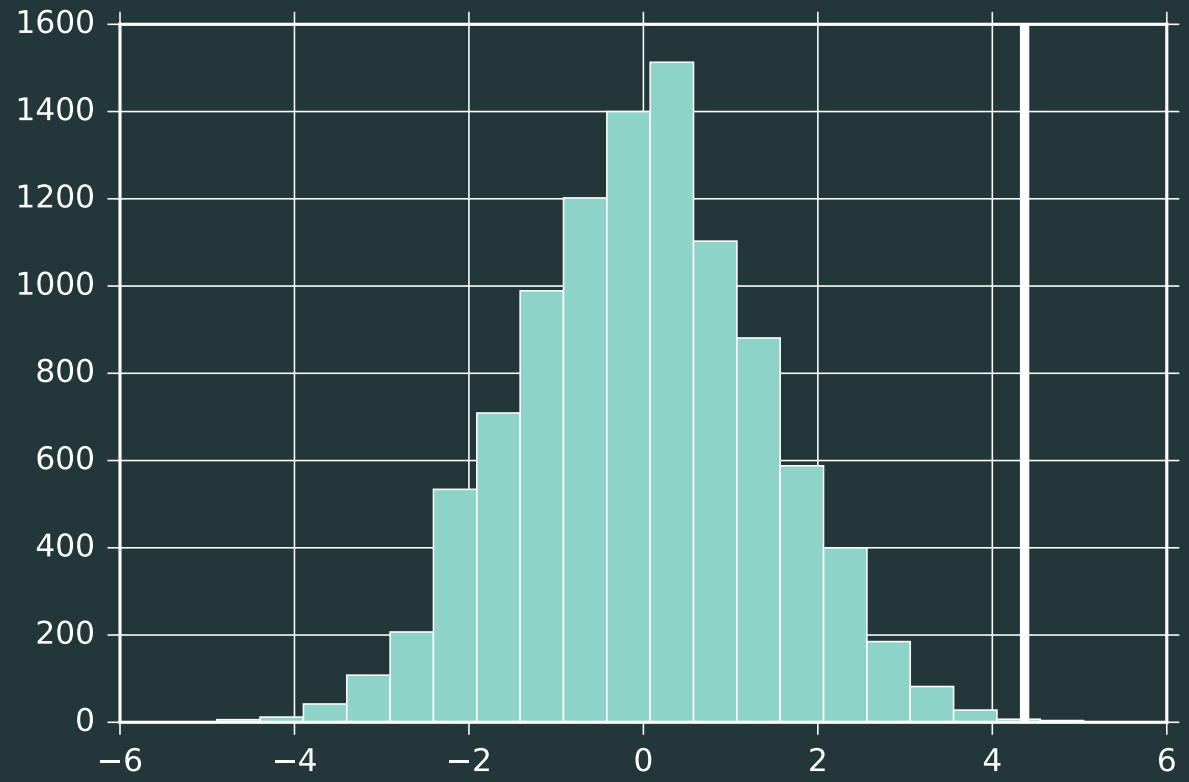
Data

1000 permutation



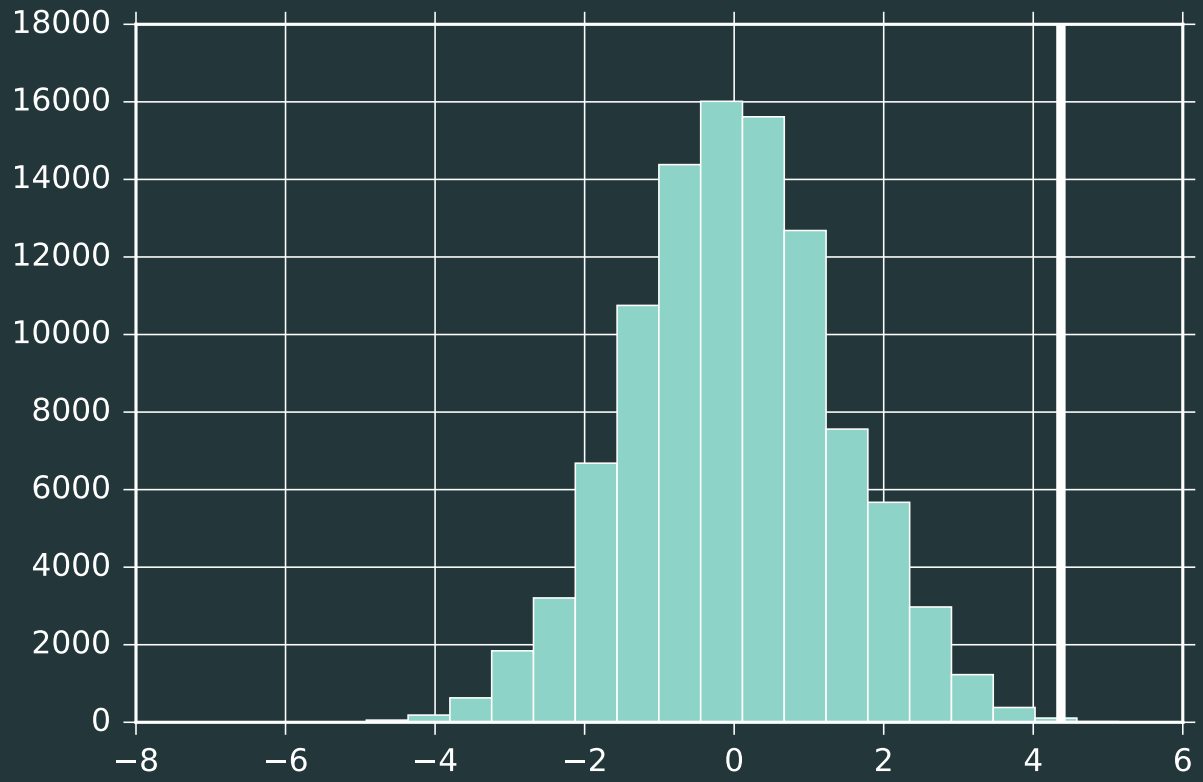
Data

10000 permutation



Data

100000 permutation



Data

We have constructed the empirical distribution of the test statistic $\text{mean}_{\text{beer}} - \text{mean}_{\text{water}}$

Data

We have constructed the empirical distribution of the test statistic $\text{mean}_{\text{beer}} - \text{mean}_{\text{water}}$

How likely is it that we arrived to a value of 4.4 by chance?

Data

We have constructed the empirical distribution of the test statistic $\text{mean}_{\text{beer}} - \text{mean}_{\text{water}}$

How likely is it that we arrived to a value of 4.4 by chance?

Easy,

$$p = \frac{\text{number of times that the statistic} \geq 4.4}{\text{total number of permutations}}$$

This is the exact definition of p-value!

In this experiment, p-value $= 0.0004$ and so the null hypothesis can be rejected.

OPEN ACCESS Freely available online

PLoS one

Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes

Thierry Lefèvre[1]*, Louis-Clément Gouagna[2,3], Kounbobr Roch Dabiré[3,4], Eric Elguero[1], Didier Fontenille[2], François Renaud[1], Carlo Costantini[2,5], Frédéric Thomas[1,6]

1 Génétique et Evolution des Maladies Infectieuses, UMR CNRS/IRD 2724, Montpellier, France, **2** Caractérisation et Contrôle des Populations de Vecteurs, IRD/UR 016, Montpellier, France, **3** Institut de Recherche en Science de la Santé, Bobo-Dioulasso, Burkina Faso, **4** Laboratoire de Parasitologie et d'Entomologie Médicale, Centre Muraz, Bobo-Dioulasso, Burkina Faso, **5** Organisation de Coordination pour la lutte contre les Endémies en Afrique Centrale, Yaoundé, Cameroun, **6** Institut de Recherche en Biologie Végétale, Université de Montréal, Montréal, Canada

Abstract

Background: Malaria and alcohol consumption both represent major public health problems. Alcohol consumption is rising in developing countries and, as efforts to manage malaria are expanded, understanding the links between malaria and alcohol consumption becomes crucial. Our aim was to ascertain the effect of beer consumption on human attractiveness to malaria mosquitoes in semi field conditions in Burkina Faso.

Now its your turn!

Go to the github repository for lecture 2

`https://github.com/dsl2l2017/lecture_2`

Do the third and last exercise.

References i


Marti Anderson and Cajo Ter Braak.
Permutation tests for multi-factorial analysis of variance.
Journal of Statistical Computation and Simulation, 2003.


Marti J Anderson.
Permutation tests for univariate or multivariate analysis of variance and regression.
Canadian journal of fisheries and aquatic sciences, 2001.


Phillip Good.
Permutation, Parametric, and Bootstrap Tests of Hypotheses.
Springer Science & Business Media, 2013.

References ii


Thierry Lefèvre, Louis-Clément Gouagna, Kounbobr Roch Dabiré, Eric Elguero, Didier Fontenille, François Renaud, Carlo Costantini, and Frédéric Thomas.
Beer consumption increases human attractiveness to malaria mosquitoes.
PloS one, 2010.