Computational Methods for Statistics (VU) (706.026)

Elisabeth Lex

ISDS, TU Graz

Jan 11, 2023

About today's class

- Motivation why computational methods are beneficial
- Inductive and deductive inference
- Sampling
- Definitions: population, random sample, sampled population, statistic
- Sampling distribution

Materials consist of slides and recommended readings.

Learning Goals

At the end of this unit, you will be able to:

- explain the benefit of using computation in statistics.
- explain sampling and why we need it.
- Define target population, random sample, sampled population, statistic.
- Explain the concept of a sampling distribution.
- Explain the difference between the distribution of a target population, the distribution of a sample, and the sampling distributions of a statistic.
- Create sampling distributions in Python.

Motivation

Doing Statistics With Computer Programs

Statistics Without the Agonizing Pain

If you can write a computer program, you have direct access to fundamental ideas in statistics.

- John Rauser (data scientist at Pinterest) keynote: `https://www.youtube.com/watch?v=5Dnw46eC-0o`

Statistics Answers Important Questions!

Research Question: Does drinking beer makes you more attractive to mosquitos?

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Beer Consumption Human Attractiveness to Malaria Mosquitoes

Thierry Lefèvre, Louis-Clément Gouagna, Kounbobr Roch Dabiré, Eric Elguero, Didier Fontenille, François Renaud, Carlo Costantini, Frédéric Thomas

Published: March 4, 2010 • https://doi.org/10.1371/journal.pone.0009546

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009546

Experiments to Collect Data

Beer

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			

Water

21	19	13	22
15	22	15	22
20	12	24	24
21	19	18	16
23	20		

$$\overline{X}_n = 23.6000 \qquad \overline{Y}_m = 19.2222$$

$$\overline{X}_n - \overline{Y}_m = 4.3778$$

Statistics Answers Important Questions!

- Average difference of ≈ 4.4 more mosquitos who were attracted to beer drinkers
- Statistical question: is this difference sufficient evidence to conclude that drinking beer makes you more attractive to mosquitos?
- We can frame this question as a debate between an advocate and a skeptic

Example 1 (Are beer drinkers more attractive to mosquitos?)

Debate between an advocate and a skeptic.

Statistics Answers Important Questions!

- Average difference of ≈ 4.4 more mosquitos who were attracted to beer drinkers
- Statistical question: is this difference sufficient evidence to conclude that drinking beer makes you more attractive to mosquitos?
- We can frame this question as a debate between an advocate and a skeptic

Example 1 (Are beer drinkers more attractive to mosquitos?)

Debate between an advocate and a skeptic.

- Advocate's argument:

Statistics Answers Important Questions!

- Average difference of ≈ 4.4 more mosquitos who were attracted to beer drinkers
- Statistical question: is this difference sufficient evidence to conclude that drinking beer makes you more attractive to mosquitos?
- We can frame this question as a debate between an advocate and a skeptic

Example 1 (Are beer drinkers more attractive to mosquitos?)

Debate between an advocate and a skeptic.

- Advocate's argument: Difference of 4.4 is large as compared to variation in the sample.

Statistics Answers Important Questions!

- Average difference of ≈ 4.4 more mosquitos who were attracted to beer drinkers
- Statistical question: is this difference sufficient evidence to conclude that drinking beer makes you more attractive to mosquitos?
- We can frame this question as a debate between an advocate and a skeptic

Example 1 (Are beer drinkers more attractive to mosquitos?)

Debate between an advocate and a skeptic.

- Advocate's argument: Difference of 4.4 is large as compared to variation in the sample.
- Skeptic's argument:

Statistics Answers Important Questions!

- Average difference of ≈ 4.4 more mosquitos who were attracted to beer drinkers
- Statistical question: is this difference sufficient evidence to conclude that drinking beer makes you more attractive to mosquitos?
- We can frame this question as a debate between an advocate and a skeptic

Example 1 (Are beer drinkers more attractive to mosquitos?)

Debate between an advocate and a skeptic.

- Advocate's argument: Difference of 4.4 is large as compared to variation in the sample.
- Skeptic's argument: Beer consumption has no effect. The difference of 4.4 is small and could have happened just by chance!

Are beer drinkers more attractive to mosquitos?

- Two approaches to settle the debate
  1. Analytical method
  2. Computational method

Are beer drinkers more attractive to mosquitos?

- Analytical method: difference of means between two samples
- Make assumptions, e.g. on the underlying distribution
- Compute the sampling distribution of the test statistic (T-statistic)
- From the sampling distribution we can compute the p-value

$$
\begin{aligned}
T &= \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}, \\
s_X &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2, s_Y = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2 \\
T &\sim f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi\Gamma\left(\frac{\nu}{2}\right)}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}
\end{aligned}
$$

Are beer drinkers more attractive to mosquitos?

- Computational method
- If the skeptic is right: the labels are meaningless

Beer

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			

Water

21	19	13	22
15	22	15	22
20	12	24	24
21	19	18	16
23	20		

$$\overline{X}_n = 23.6000$$

$$\overline{Y}_m = 19.2222$$

$$\overline{X}_n - \overline{Y}_m = 4.3778$$

Are beer drinkers more attractive to mosquitos?

Beer

16	13	20	19
23	26	20	21
19	22	20	24
20	21	20	23
21	24	25	27
15	22	24	21
15			

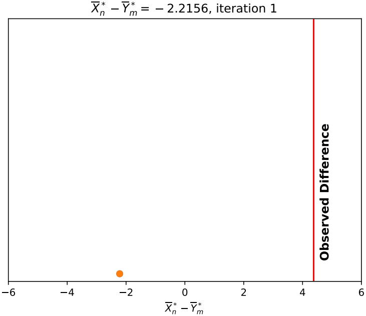
Water

28	27	18	31
21	29	22	24
27	12	19	17
24	19	28	31
20	18		

$\overline{X}_n^* = 20.8400$

$\overline{Y}_m^* = 23.0556$

$$\overline{X}_n^* - \overline{Y}_m^* = -2.2156$$



Are beer drinkers more attractive to mosquitos?

Beer

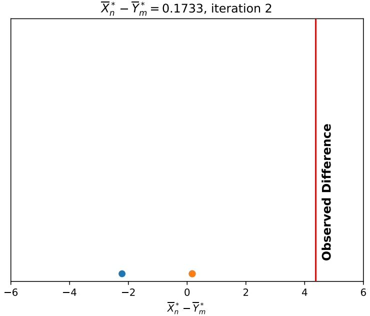
19	27	20	17
22	26	20	24
23	19	27	20
31	19	20	18
18	28	15	21
27	29	13	19
24			

Water

24	21	28	23
22	21	21	15
12	24	16	21
31	20	24	22
20	25		

$\overline{X}_n^* = 21.8400$ $\overline{Y}_m^* = 21.6667$

$$\overline{X}_n^* - \overline{Y}_m^* = 0.1733$$



Are beer drinkers more attractive to mosquitos?

Beer

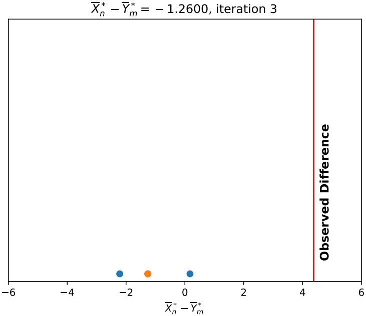
21	18	20	16
19	20	15	21
24	20	21	24
17	21	31	22
22	12	28	28
20	19	23	25
24			

Water

18	24	29	31
20	22	24	19
15	27	20	27
21	19	27	13
26	23		

$$\overline{X}_n^* = 21.2400 \qquad \overline{Y}_m^* = 22.5000$$

$$\overline{X}_n^* - \overline{Y}_m^* = -1.2600$$



Are beer drinkers more attractive to mosquitos?

Beer

25	26	16	28
31	22	19	27
20	20	21	31
22	27	22	24
20	13	21	21
20	27	18	23
24			

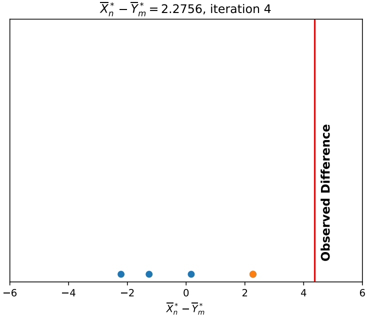
Water

12	19	18	23
20	19	20	15
24	21	15	24
29	19	28	21
24	17		

$\overline{X}_n^* = 22.7200$

$\overline{Y}_m^* = 20.4444$

$\overline{X}_n^* - \overline{Y}_m^* = 2.2756$



Are beer drinkers more attractive to mosquitos?

Beer

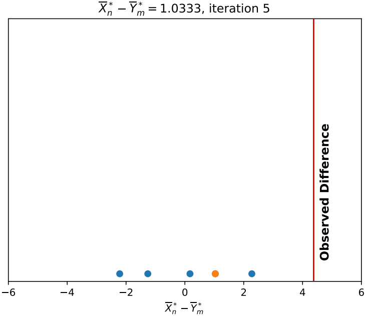
31	19	19	15
23	21	24	21
22	25	28	22
20	19	29	27
20	20	24	15
21	27	23	28
12			

Water

24	20	22	27
24	21	21	13
19	20	24	16
26	17	18	20
31	18		

$\overline{X}_n^* = 22.2000$ $\overline{Y}_m^* = 21.1667$

$$\overline{X}_n^* - \overline{Y}_m^* = 1.0333$$



Are beer drinkers more attractive to mosquitos?

Beer

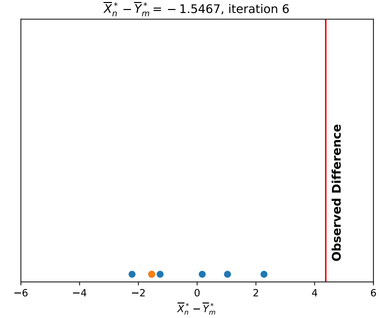
21	23	23	28
21	19	29	24
16	19	28	20
18	27	21	20
15	19	24	12
15	22	24	19
21			

Water

20	27	13	21
20	24	24	31
22	25	26	31
18	20	20	17
22	27		

$\overline{X}_n^* = 21.1200$ $\overline{Y}_m^* = 22.6667$

$$\overline{X}_n^* - \overline{Y}_m^* = -1.5467$$



Are beer drinkers more attractive to mosquitos?

Beer

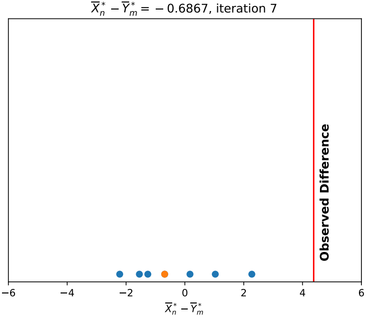
20	20	21	21
16	18	26	23
24	28	15	21
13	15	20	27
28	22	19	20
24	29	23	24
20			

Water

31	24	22	27
19	25	22	21
12	19	21	27
31	20	17	18
19	24		

$\overline{X}_n^* = 21.4800$ $\overline{Y}_m^* = 22.1667$

$$\overline{X}_n^* - \overline{Y}_m^* = -0.6867$$



Are beer drinkers more attractive to mosquitos?

Beer

20	28	19	22
21	15	23	31
29	24	27	19
19	27	26	18
28	27	20	20
19	16	24	17
31			

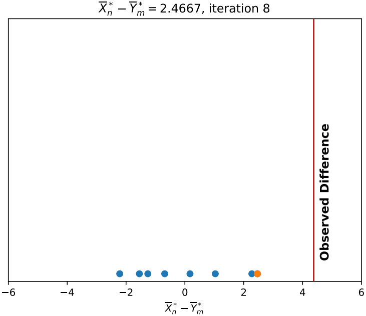
Water

20	21	21	22
25	12	24	18
24	15	21	21
13	22	24	20
20	23		

$\overline{X}_n^* = 22.8000$

$\overline{Y}_m^* = 20.3333$

$\overline{X}_n^* - \overline{Y}_m^* = 2.4667$



Are beer drinkers more attractive to mosquitos?

Beer

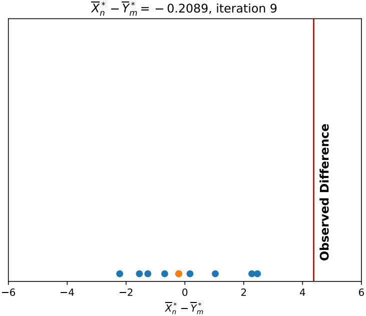
28	24	20	19
21	24	23	15
29	19	20	20
25	18	19	24
19	27	20	15
26	24	20	22
21			

Water

12	17	13	16
22	23	20	27
21	21	31	21
22	31	18	24
28	27		

$\overline{X}_n^* = 21.6800$ $\quad\overline{Y}_m^* = 21.8889$

$$\overline{X}_n^* - \overline{Y}_m^* = -0.2089$$



Are beer drinkers more attractive to mosquitos?

Beer

23	17	20	20
28	26	24	29
19	19	16	18
22	27	15	19
21	31	21	12
25	20	21	20
27			

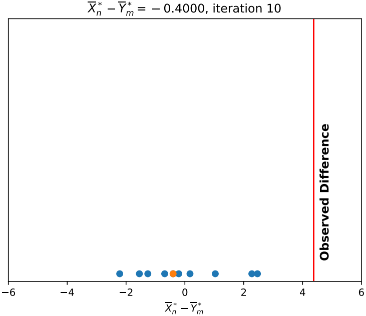
$\overline{X}_n^* = 21.6000$

Water

21	20	22	22
15	21	24	18
31	24	27	13
19	20	23	24
24	28		

$\overline{Y}_m^* = 22.0000$

$$\overline{X}_n^* - \overline{Y}_m^* = -0.4000$$



Are beer drinkers more attractive to mosquitos?

Beer

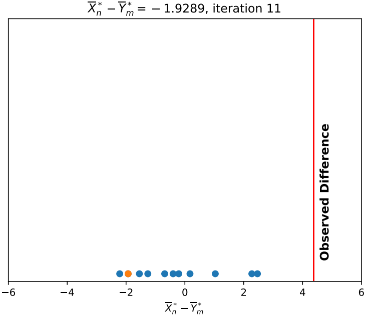
21	24	21	16
20	19	21	24
20	29	19	20
13	25	31	24
21	15	18	19
18	22	22	20
22			

Water

19	21	28	24
20	20	12	28
23	17	15	26
31	24	27	27
27	23		

$\overline{X}_n^* = 20.9600$ $\qquad$ $\overline{Y}_m^* = 22.8889$

$$\overline{X}_n^* - \overline{Y}_m^* = -1.9289$$



Are beer drinkers more attractive to mosquitos?

Beer

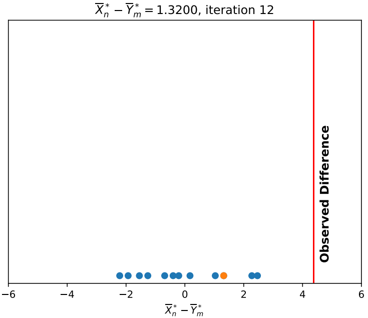
21	23	21	27
24	22	29	20
24	15	18	21
24	20	25	22
19	20	20	19
31	20	20	27
26			

Water

19	27	22	13
19	18	28	24
21	24	21	17
28	23	15	16
12	31		

$\overline{X}_n^* = 22.3200$ $\quad$ $\overline{Y}_m^* = 21.0000$

$$\overline{X}_n^* - \overline{Y}_m^* = 1.3200$$



Are beer drinkers more attractive to mosquitos?

Beer

24	15	21	22
20	31	17	28
24	18	20	21
21	20	27	13
18	31	12	26
27	21	23	22
24			

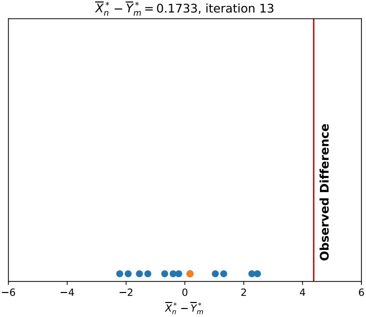
Water

19	20	20	29
19	23	28	24
20	16	22	24
19	25	27	15
19	21		

$\overline{X}_n^* = 21.8400$

$\overline{Y}_m^* = 21.6667$

$$\overline{X}_n^* - \overline{Y}_m^* = 0.1733$$



Are beer drinkers more attractive to mosquitos?

Beer

20	21	27	28
24	27	31	19
21	12	19	19
28	13	22	24
23	25	21	20
23	24	22	19
20			

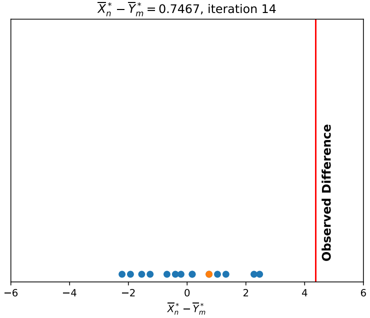
$\overline{X}_n^* = 22.0800$

Water

24	16	22	17
15	20	20	29
18	21	31	15
18	26	21	27
20	24		

$\overline{Y}_m^* = 21.3333$

$$\overline{X}_n^* - \overline{Y}_m^* = 0.7467$$



Are beer drinkers more attractive to mosquitos?

Beer

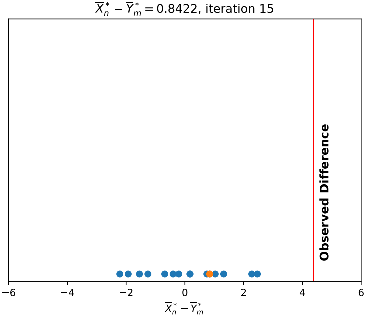
20	16	17	28
19	27	24	20
31	26	23	22
29	22	19	25
22	13	21	21
18	24	28	20
18			

Water

31	21	21	27
24	21	27	15
20	19	20	19
20	24	12	23
15	24		

$\overline{X}_n^* = 22.1200$ $\overline{Y}_m^* = 21.2778$

$$\overline{X}_n^* - \overline{Y}_m^* = 0.8422$$



Are beer drinkers more attractive to mosquitos?

Beer

18	17	20	31
28	19	20	25
26	15	28	13
20	21	18	22
21	15	24	22
21	27	23	23
22			

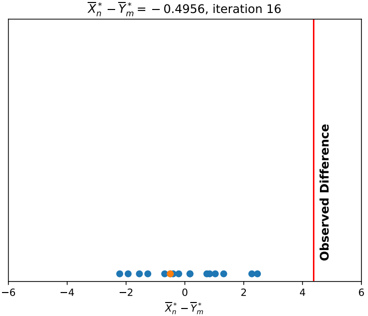
Water

19	21	24	20
19	24	12	21
20	27	31	27
20	29	24	24
19	16		

$\overline{X}_n^* = 21.5600$

$\overline{Y}_m^* = 22.0556$

$$\overline{X}_n^* - \overline{Y}_m^* = -0.4956$$



Are beer drinkers more attractive to mosquitos?

Beer

24	20	15	24
27	22	21	20
28	28	21	12
21	18	25	21
29	18	31	24
17	22	13	16
23			

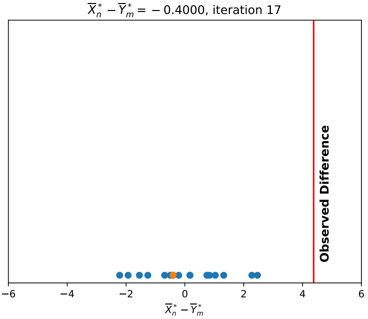
Water

19	31	20	20
19	15	24	22
20	19	20	21
24	19	27	27
26	23		

$\overline{X}_n^* = 21.6000$

$\overline{Y}_m^* = 22.0000$

$$\overline{X}_n^* - \overline{Y}_m^* = -0.4000$$



Are beer drinkers more attractive to mosquitos?

Beer

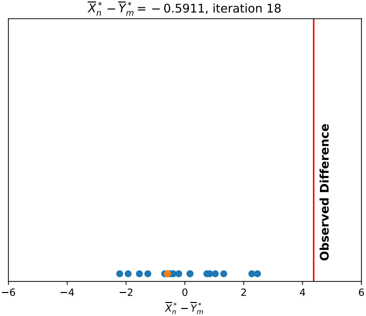
15	24	15	28
19	16	22	21
24	26	18	20
20	27	19	19
21	20	21	18
20	31	21	25
28			

Water

17	31	22	20
23	23	12	27
24	21	22	13
24	20	19	29
24	27		

$\overline{X}_n^* = 21.5200$ $\quad$ $\overline{Y}_m^* = 22.1111$

$$\overline{X}_n^* - \overline{Y}_m^* = -0.5911$$



Are beer drinkers more attractive to mosquitos?

Beer

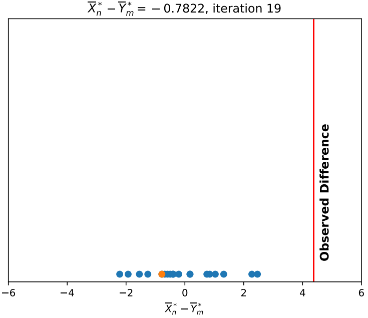
22	15	31	22
23	19	21	12
18	20	27	20
24	27	23	25
27	22	24	17
20	19	24	13
21			

Water

28	18	20	16
15	21	19	24
28	20	31	21
21	19	20	29
24	26		

$\overline{X}_n^* = 21.4400$ $\overline{Y}_m^* = 22.2222$

$$\overline{X}_n^* - \overline{Y}_m^* = -0.7822$$



Are beer drinkers more attractive to mosquitos?

Beer

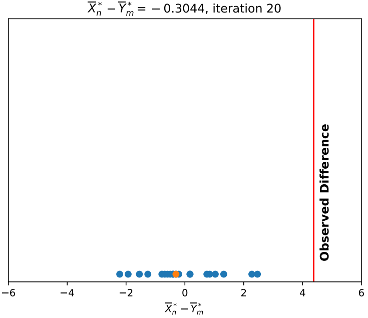
20	21	19	23
18	24	23	19
12	17	29	18
28	22	15	13
21	27	31	24
21	19	22	31
24			

Water

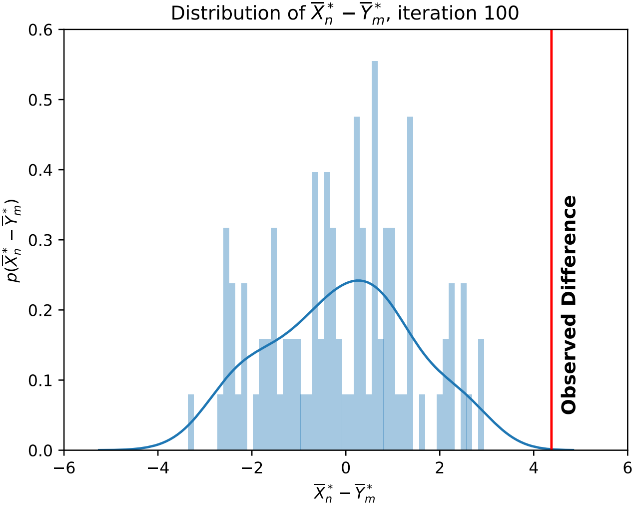
19	20	24	28
20	26	27	24
16	15	25	21
21	20	22	20
27	20		

$\overline{X}_n^* = 21.6400$ $\quad$ $\overline{Y}_m^* = 21.9444$

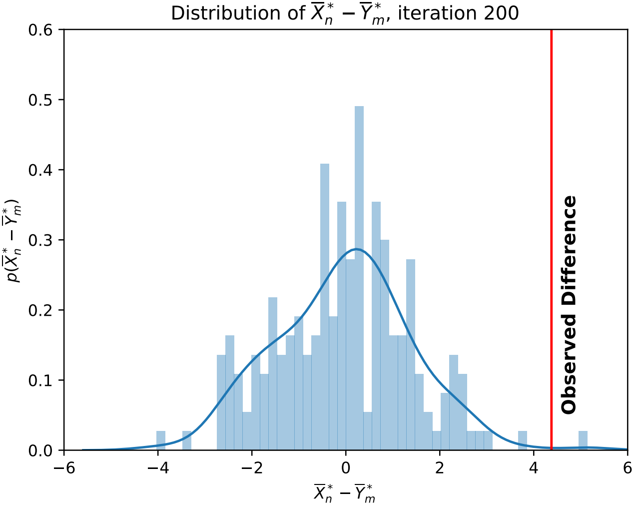
$$\overline{X}_n^* - \overline{Y}_m^* = -0.3044$$



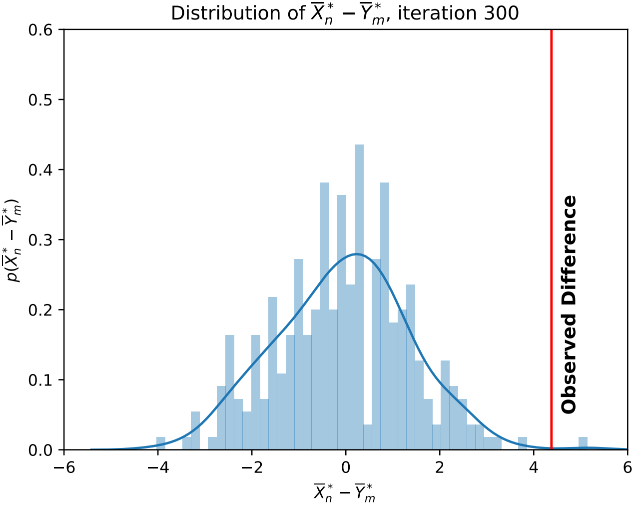
Are beer drinkers more attractive to mosquitos?



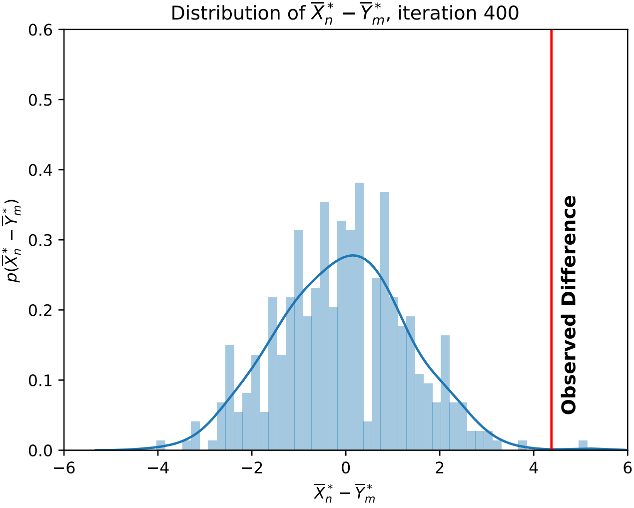
Are beer drinkers more attractive to mosquitos?



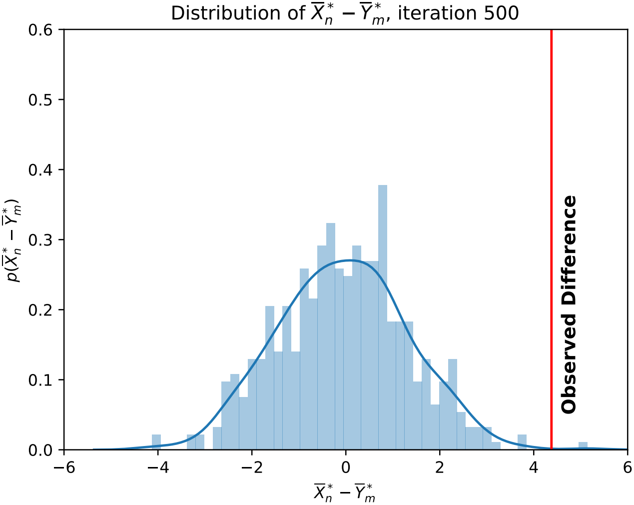
Are beer drinkers more attractive to mosquitos?



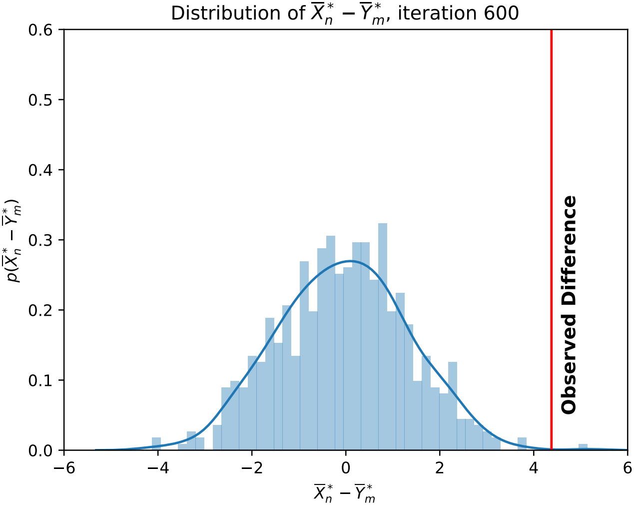
Are beer drinkers more attractive to mosquitos?



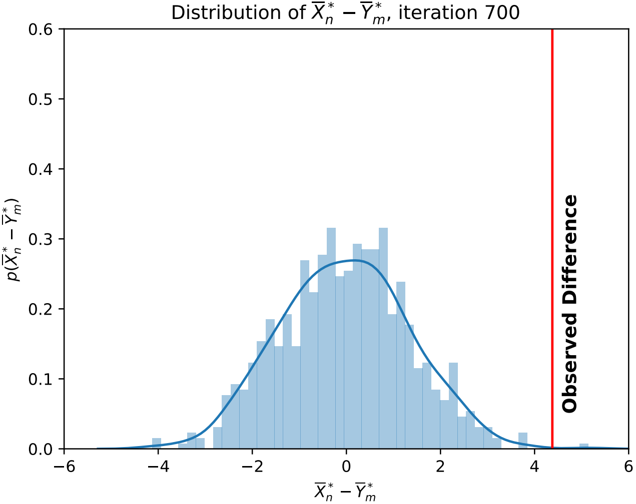
Are beer drinkers more attractive to mosquitos?



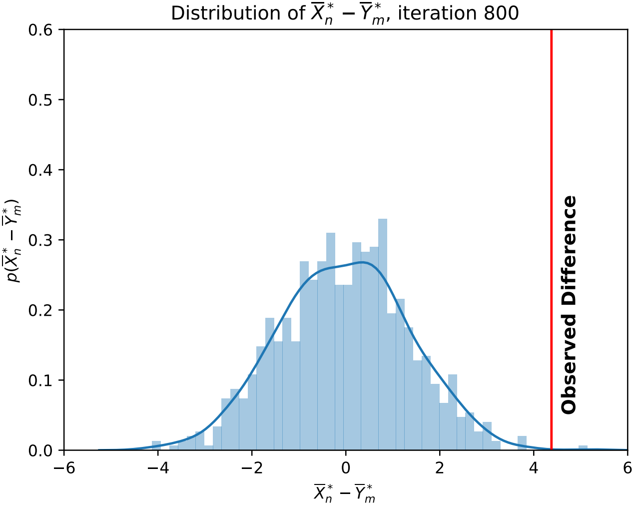
Are beer drinkers more attractive to mosquitos?



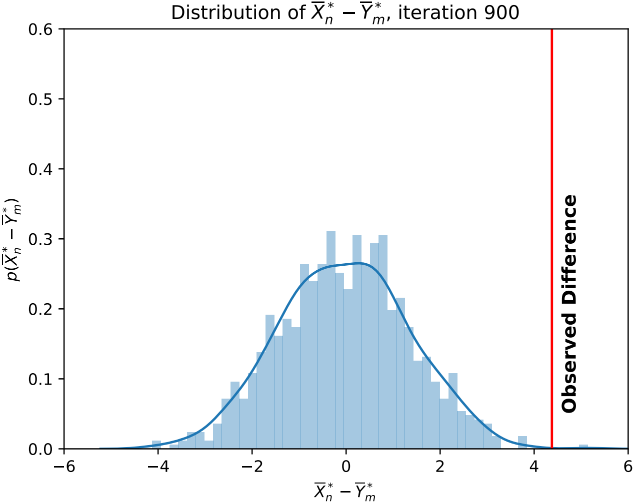
Are beer drinkers more attractive to mosquitos?



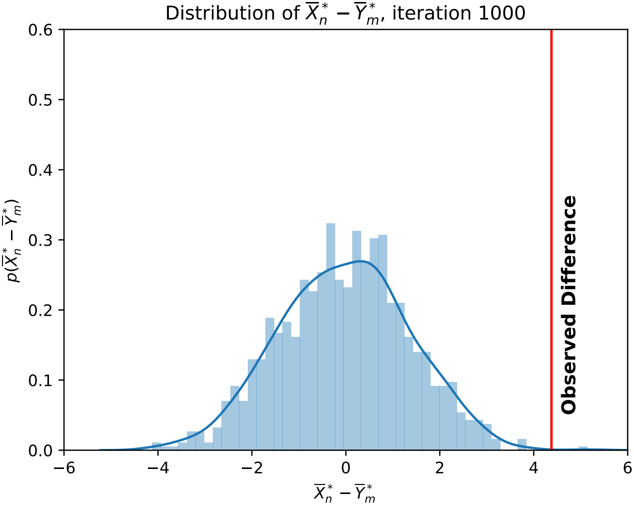
Are beer drinkers more attractive to mosquitos?



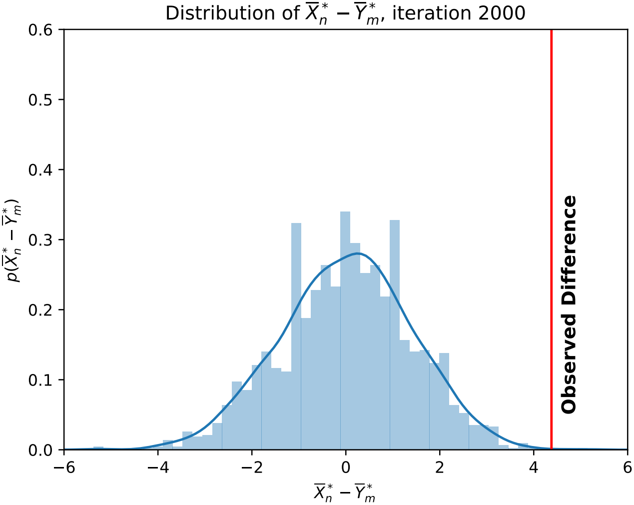
Are beer drinkers more attractive to mosquitos?



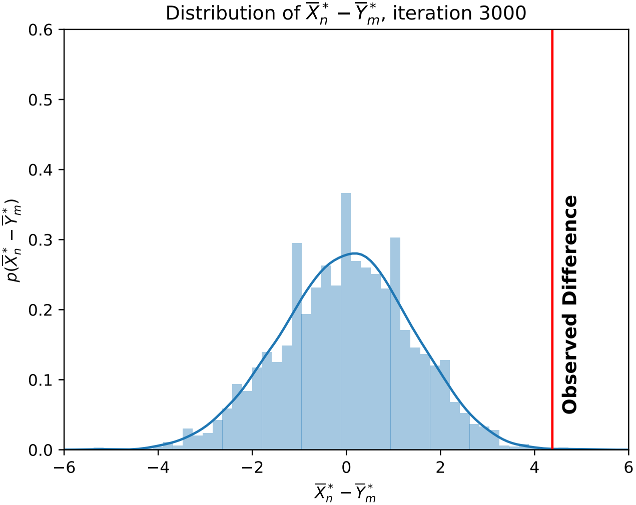
Are beer drinkers more attractive to mosquitos?



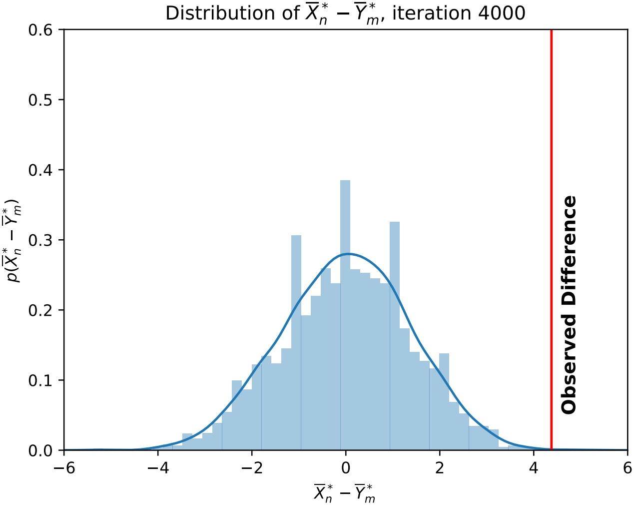
Are beer drinkers more attractive to mosquitos?



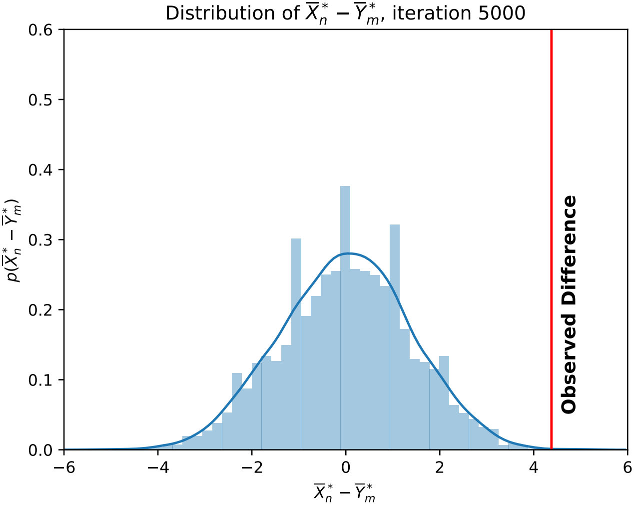
Are beer drinkers more attractive to mosquitos?



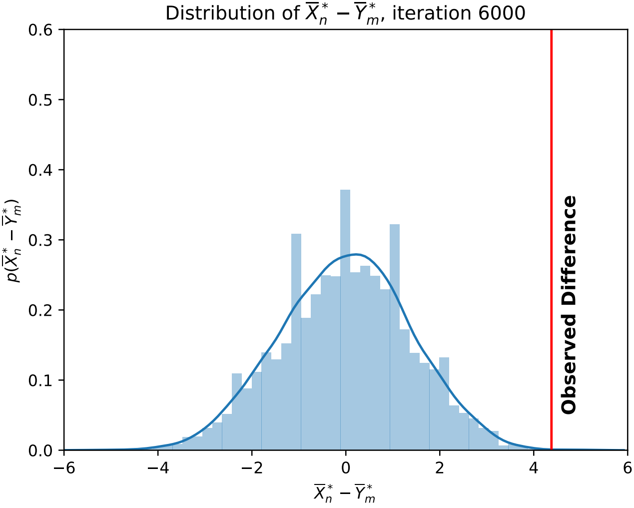
Are beer drinkers more attractive to mosquitos?



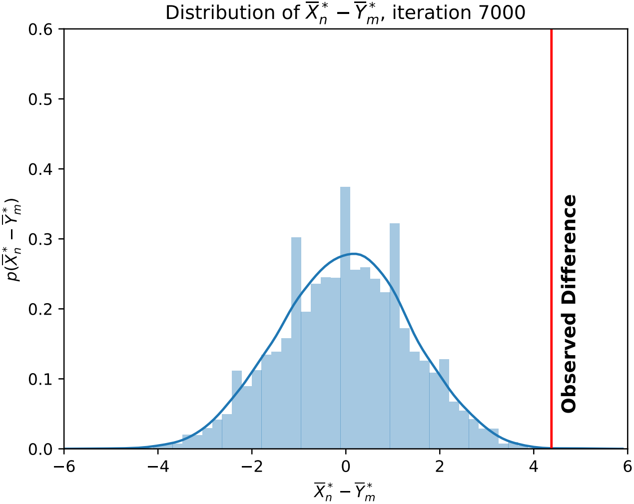
Are beer drinkers more attractive to mosquitos?



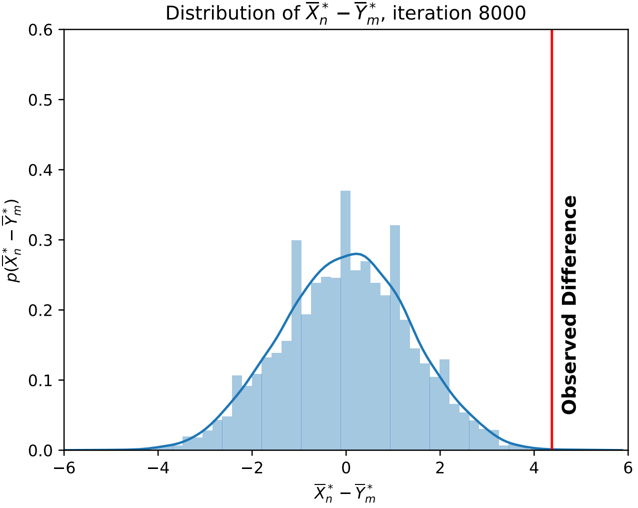
Are beer drinkers more attractive to mosquitos?



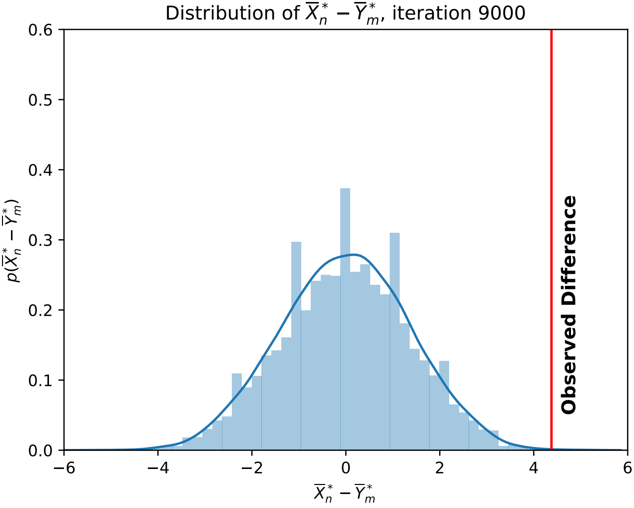
Are beer drinkers more attractive to mosquitos?



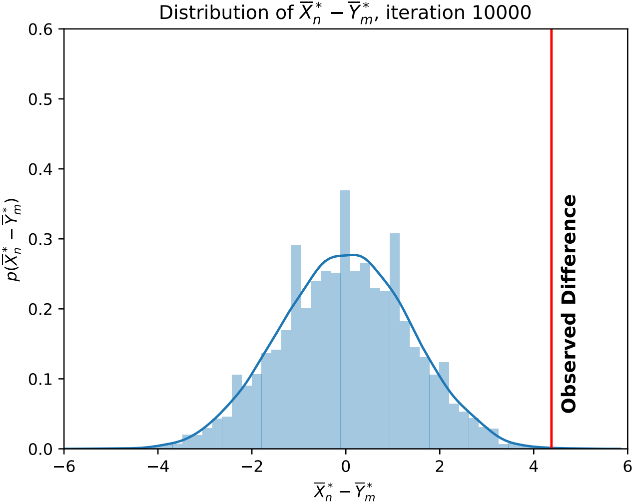
Are beer drinkers more attractive to mosquitos?



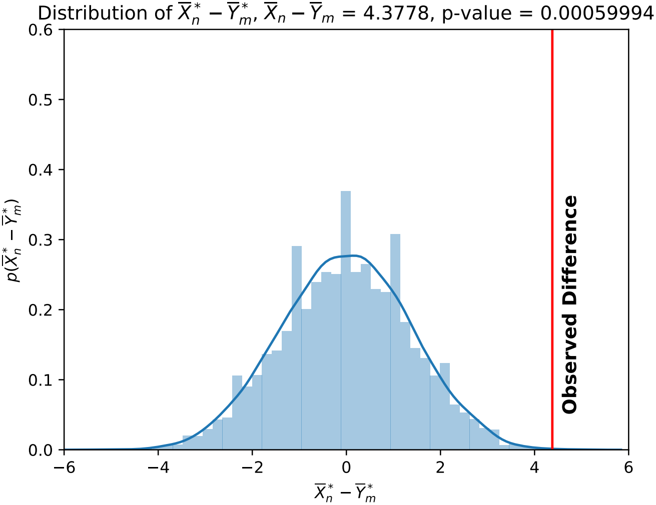
Are beer drinkers more attractive to mosquitos?



Are beer drinkers more attractive to mosquitos?



Are beer drinkers more attractive to mosquitos?



Are beer drinkers more attractive to mosquitos?

Research Question :)

Does drinking beer makes you more attractive to mosquitos? Yes, it does!

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes

Thierry Lefèvre, Louis-Clément Gouagna, Kounbobr Roch Dabiré, Eric Elguero, Didier Fontenille, François Renaud, Carlo Costantini, Frédéric Thomas

Published: March 4, 2010 • https://doi.org/10.1371/journal.pone.0009546

241 Save	30 Citation
49,472 View	327 Share

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009546

Computational Method

- Computational method: random permutation test
- We will see more examples of this method in the course
- Resampling methods: resample and iterate
- **Key message: if you can program, you can play with statistics**

Inductive Inference

Readings

Fahrmeir et al., Statistik: der Weg zur Datenanalyse (Chapters 1.4, 7)

Wasserman, All of Statistics (Chapter 5)

Inductive Inference

- Scientific progress is strongly related to experimentation.
- Experiments result in data.
- We then draw conclusions from the data.
- Frequently, we generalize from a single or from a few experiments.
- **Inductive inference is uncertain.**
- Statistics provides principles for measuring that uncertainty.
- We measure uncertainty with probabilities.

Deductive Inference

- Inductive inference results in probabilities.
- Deductive inference is conclusive:

Example 2 (Right triangles)

Two statements:

1. One of the angles of each right triangle is 90°.
2. Triangle A is a right triangle.

Conclusion: One of the angles of triangle A is 90°.

Inductive Inference

- Deductive inference: prove theorems in mathematics
- Inductive inference: find new knowledge in empirical research

Example 3 (Rise of Skywalker)

Suppose we want to know what percent of the Austrian population watched the Star Wars movie. The only way to answer this exactly is to ask all Austrians if they watched the movie, which is not feasible. Thus, we ask a few Austrians about the movie, and on the basis of their responses we make probabilistic statements or predictions for the whole population.

Target Population

- We observe a few of the elements of interest.
- Based on these, we make a statement about the totality of elements.

Definition 1 (Target population)

We call the totality of elements of interest *target population*.

Example 3 (Rise of Skywalker)

All people living in Austria constitute the target population.

Methodology for inductive inference

- Goal: find out something about a certain target population.
- Impractical - and often impossible - to examine the entire population.
- Examine a **sample** (a part) of it.
- Make inferences regarding the entire target population.

Sampling

Random Sample

Definition 2 (Random sample)

Let the random variables $X_1, X_2, \ldots, X_n$ have a joint density $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ that factors as follows:

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdot \cdots \cdot f(x_n),$$

where $f(\cdot)$ is the common density of each X_i. We then define $X_1, X_2, \ldots, X_n$ to be a *random sample* of size n from a population with density $f(\cdot)$. Thus, a random sample is a sequence of independent, identically distributed (*i.i.d.*) random variables.

Remark 1 (Sampling with/without replacement)

When sampling from a finite population, our definition requires to always sample with replacement as otherwise the drawings are not independent.

Random Sample

Example 3 (Rise of Skywalker)

We define X_i as 1 (ith person watched the movie) or as 0 (ith person did not watch the movie). If we sample people so that the variables $X_1, X_2, \ldots, X_n$ are independent and have the same density (all people have the same probability of watching the movie) then the sample is random.

Sampled Population

Definition 3 (Sampled population)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density $f(\cdot)$; then this population is called *sampled population*.

Remark 2 (Distinction between the sampled and the target population)

- With random samples we can only make valid probability statements about sampled population.
- Statements about target population are not valid.
- Unless the target population is also the sampled population.

Example 3 (Rise of Skywalker)

All people living in Austria form the *target population*. We draw a sample from Graz. Thus, Graz residents form the *sampled population*.

Methodology for inductive inference revisited

- Goal: study a population with density $f(\cdot;\theta)$.
- We know the form of the density, but it contains an unknown parameter θ.
- Take a random sample $X_1, X_2, \ldots, X_n$ of size n from $f(\cdot;\theta)$.
- We compute the value of some function $t(x_1, x_2, \ldots, x_n)$ to estimate θ.

Remark 3 (Terminology clarification)

- $X_i's$ are random variables (mathematical objects).
- $x_i's$ are realizations or data points (concrete observations).

Statistic

- The function, which we compute on some concrete realizations, is called **statistic**.

Definition 4 (Statistic)

Let $X_1, X_2, \dots, X_n$ be a random sample of size n from density $f(\cdot)$. Statistic is a function $T_n = t(X_1, X_2, \dots, X_n)$.

Remark 4 (Properties of statistics)

- A function of a random variable is also a random variable. $\implies$ each statistic is a random variable.
- Each random variable has a density. $\implies$ each statistic has a density.
- For probabilistic properties, we study *sampling distribution* of T_n.
- For a concrete application, we study $t_n = t(x_1, x_2, \dots, x_n)$.

Statistic

- Statistic cannot depend on unknown parameters θ if we have a density $f(\cdot;\theta)$.
- For example, if the random variable X has a density $f(\cdot;\mu,\sigma^2)$, where μ and σ^2 are unknown, then:
  - $X-\mu$ is not a statistic.
  - $\frac{X}{\sigma}$ is not a statistic.
  - X, $X+4$, X^2, X^3+lnX^2 are all statistics.

Remark 5 (Rule of thumb)

If you can write a computer function to compute a value only from your data, then it is a statistic.

Example: Sample Mean (is a Statistic)

Definition 5 (Sample mean)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the density $f(\cdot)$. *Sample mean* is defined as the average value:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

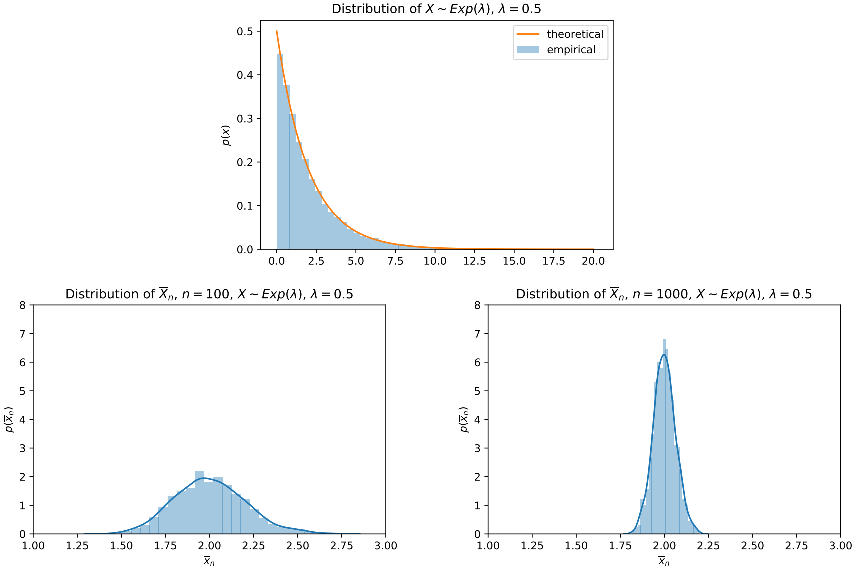
Example 4 (Sampling distribution of the sample mean)

Let $X_1, X_2, \ldots, X_n$ be a random sample with $X_i \sim Exp(\lambda)$. Estimate the sampling distribution of $\overline{X}_n$ by repeatedly drawing random numbers in python with $n = 100$ and $\lambda = 0.5$. Repeat the experiment for $n = 1000$.

Notebook 1 (Sample mean)

`sample_mean.ipynb`

Sampling Distribution of $\overline{X}_n$



Summary

- Computational methods are extremely useful to efficiently do statistics.
- Sampling: we do not investigate the whole population, but we take a sample, compute a statistic related to the parameter of interest and make an inference.
- We defined: random sample, sampled population, statistic, sampling distribution.
- Sampling distribution of the statistic tells us how close the statistic is to the parameter.

Thank you for your attention - Questions?

Next time: sampling continued, inference