# **Statistical Inference**

🎯 How can we guess the real value of a parameter based *only* on a limited sample of observations ?

1. Collect some observations of a parameter
2. Infer the true value of the parameter (leap of faith)
3. Estimate your level of confidence

### **Plan**

1. Motivation
2. Probability Theory reminders
3. Sampling Distribution and Confidence Intervals
4. Hypothesis Testing (p-values)
5. t-tests
6. Bayesian Inference

## **1. Motivation**

#### **Recall our business problem**

How to increase customer satisfaction while maintaining a healthy order volume?

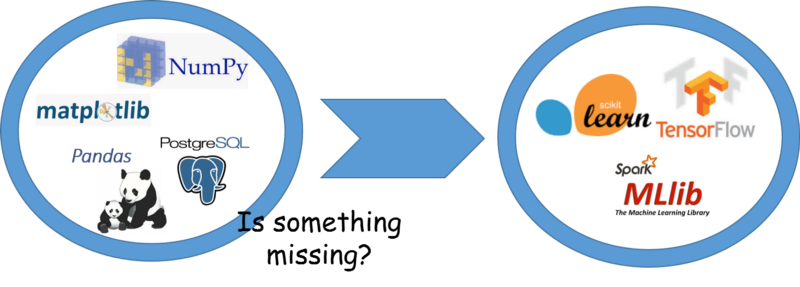
The customer satisfaction can be evaluted through the review\_score

👉 We will investigate which features are the most impactful on review\_score

Imagine we find wait\_time to be strongly correlated with bad review\_score

🤔 How can we be **confident** our findings (on historical orders) will **generalize** well ?

❌ We cannot wait years to prove that our findings were right or wrong!



**Welcome Statistical Inference Analysis!**

1️⃣ Train *linear ML* models to find correlations

2️⃣ Use stats (Central Limit Theorem!) to **quantify the statistical significance** of our findings

## **2. Probability Refreshers from the Maths module**

**Probability**Conditional Probability

P

(

B

|

A

)

Bayes Theorem

P

(

B

|

A

)

=

P

(

A

|

B

)

P

(

B

)

P

(

A

)

**Random variable**

X

= numerical outcome of a random experiment

**Random process**

X

=

(

X

k

)

0

≤

k

≤

n

= repeated sequence of random experiments

**Probability Distribution**

p

(

X

)

=

p

(

μ

,

σ

,

.

.

.

)

* Measures the underlying distribution of a random variable X
* The *mean*
* μ
* and *standard deviation*
* σ
* are called "statistics" that "describe" X
* Other statistics include kurtosis etc...

**The Gaussian Distribution (or Normal Distribution)**

N

(

μ

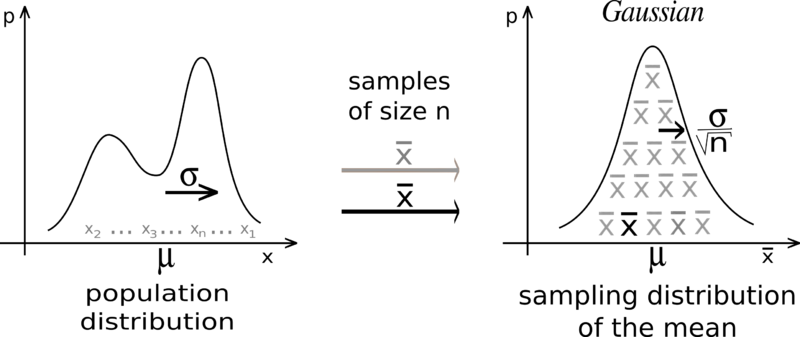
,

σ

)

* is completely described by these two statistics only

**Central Limit Theorem**

****

When you consider **independent random variables**

X

1

…

X

n

with a **common** underlying probability distribution

p

(

μ

,

σ

)

:

* Their mean
* ¯¯¯¯¯
* X
* converges towards a Normal Distribution as
* n
* increases:
  + centered around the common mean
  + μ
  + ¯¯¯¯¯
  + X
  + =
  + μ
  + with a standard deviation
  + σ
  + ¯¯¯¯¯
  + X
  + =
  + σ
  + √
  + n

¯¯¯¯¯

X

=

X

1

+

.

.

.

+

X

n

n

≈

n

→

∞

N

(

μ

,

σ

√

n

)

#### **z-score**

* If
* x
* is an observation derived from a random variable
* X
* (
* μ
* ,
* σ
* )
* , we define its z-score as follows:

z

=

x

−

μ

σ

* z = value of
* x
* expressed in *number of standard deviations above/below the mean*
* μ

Z

=

(

¯¯¯¯¯

X

−

μ

σ

√

n

)

−−−→

n

→

∞

N

(

0

,

1

)

## **3. Sampling Distribution**

### **🥋 How to estimate the average height of US citizens ?**

🎯 If my goal is to estimate the average height

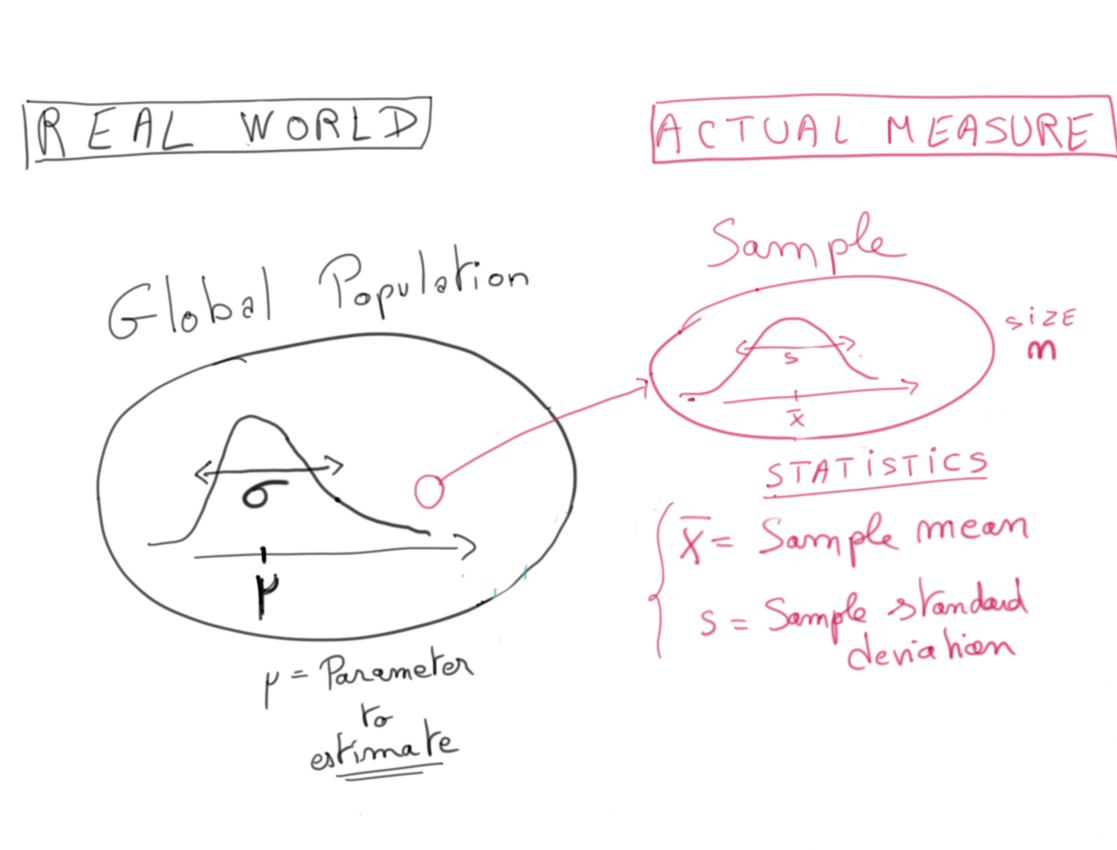
μ

among the US citizens:

❌ I can't measure the entire US population (N = 331 M)

#### **Random sampling method**

* I randomly select a sample of size n = 1000 people from the population 🎲
* Based on these 1000 people, I can compute 🧮 :
  + the sample mean
  + ¯¯¯¯¯¯¯
  + X
  + n
  + = 170 cm
  + the sample standard deviation
  + s
  + = 20 cm



#### **❓ What does it say about**

#### μ

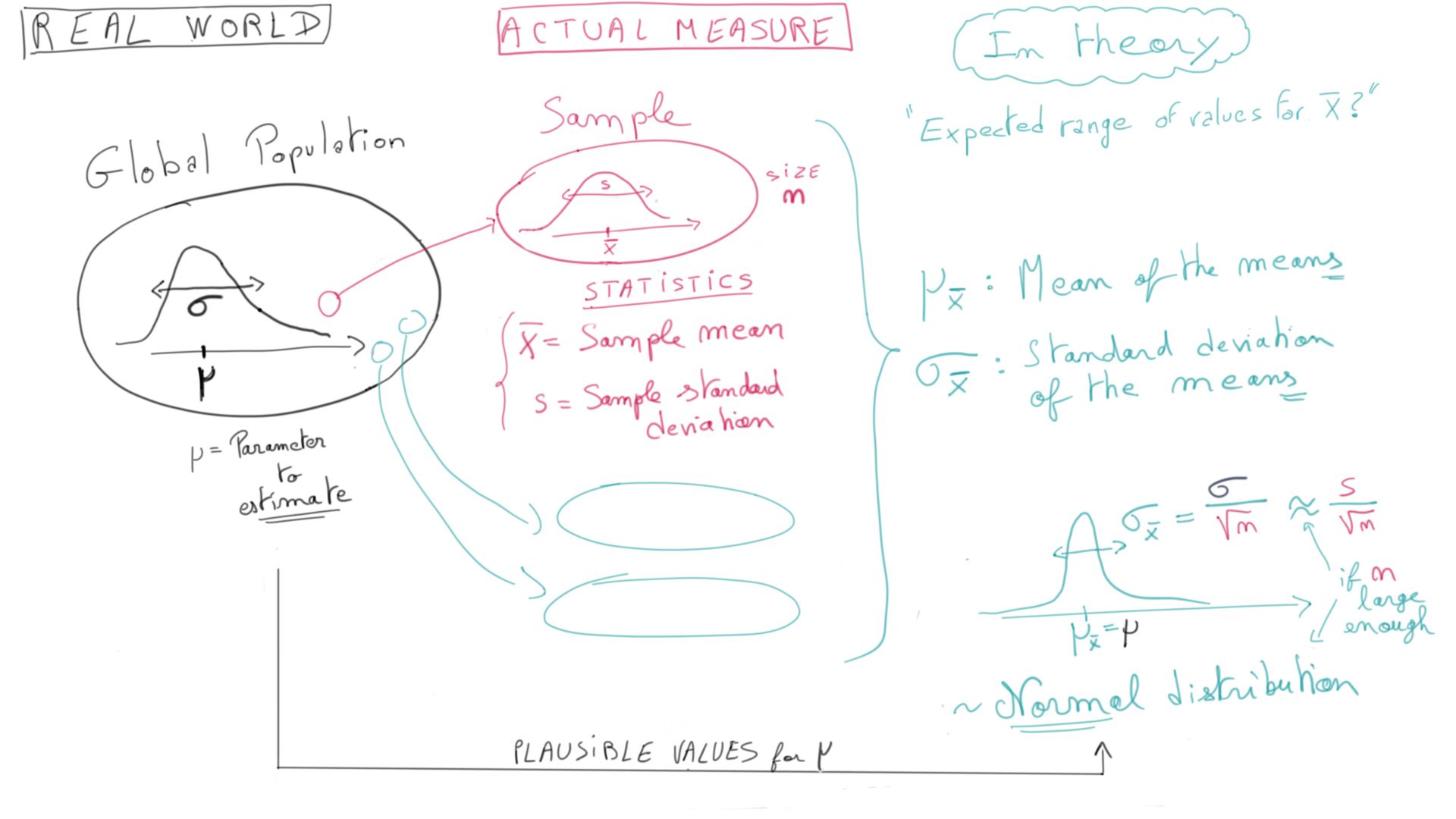
#### **❓**

**Best Guess**

* Our best estimation for
* μ
* is
* ¯¯¯¯¯¯¯
* X
* n
* = 170 cm
* This intuitive fact is due to the Law of Large Numbers:   
  When you consider **independent random variables**
* X
* 1
* …
* X
* n
* with a **common** underlying probability distribution
* p
* (
* μ
* ,
* σ
* )
* , their average
* ¯¯¯¯¯¯¯
* X
* n
* becomes a strong approximation of
* μ
* as the sample size
* n
* increases:
* ¯¯¯¯¯¯¯
* X
* n
* =
* X
* 1
* +
* .
* .
* .
* +
* X
* n
* n
* −−−→
* n
* →
* ∞
* μ

**Confidence Interval**

* We can also give a *distribution of plausible values* for
* μ
* 🎉
* Thanks to the Central Limit Theorem



Because

n

is large enough, and the citizens are randomly selected (CLT):

The distribution of sample mean**s**

¯¯¯¯¯¯¯

X

n

should follow the normal distribution:

¯¯¯¯¯¯¯

X

n

≈

N

(

μ

,

σ

√

n

)

≈

N

(

μ

,

s

√

n

)

=

N

(

μ

,

20

√

1000

)

=

N

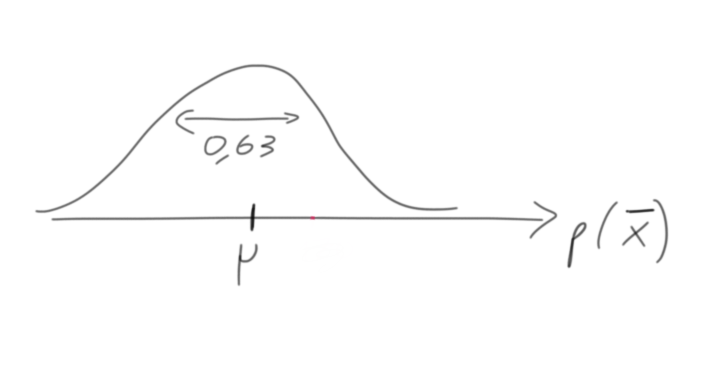
(

μ

,

0.63

)



* So we know that
* ¯¯¯¯¯¯¯
* X
* n
* *should* be centered round
* N
* (
* μ
* ,
* 0.63
* )
* And yet we *did* measure
* ¯¯¯¯¯¯¯
* X
* n
* =
* 170
* cm
* What distribution for
* μ
* is therefore the **most plausible / likely**?

N

(

170

,

0.63

)

We say that

¯¯¯¯¯¯¯

X

n

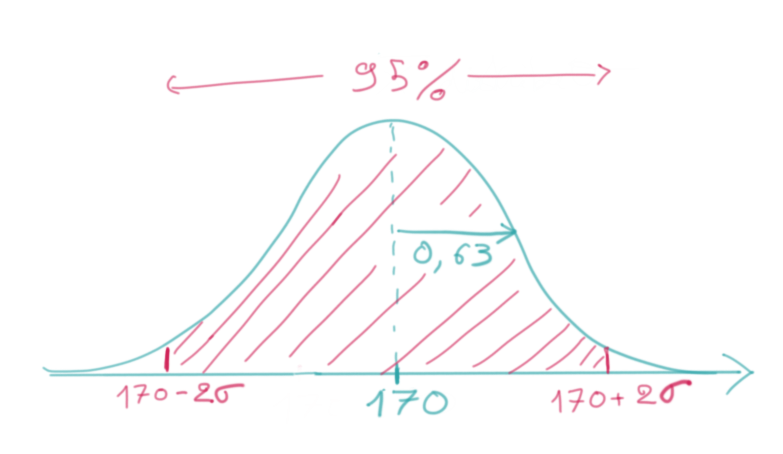
= 170 cm is the "**Maximum Likelihood Estimate (MLE)**" for

μ

Estimated probability for

μ

:



👉 We read :

μ

=

170

±

2

×

0.63

[95% confidence interval]

⇔

μ

=

170

±

1.26

c

m

[95% confidence interval]

⇔

μ

is between 168.7 and 171.2

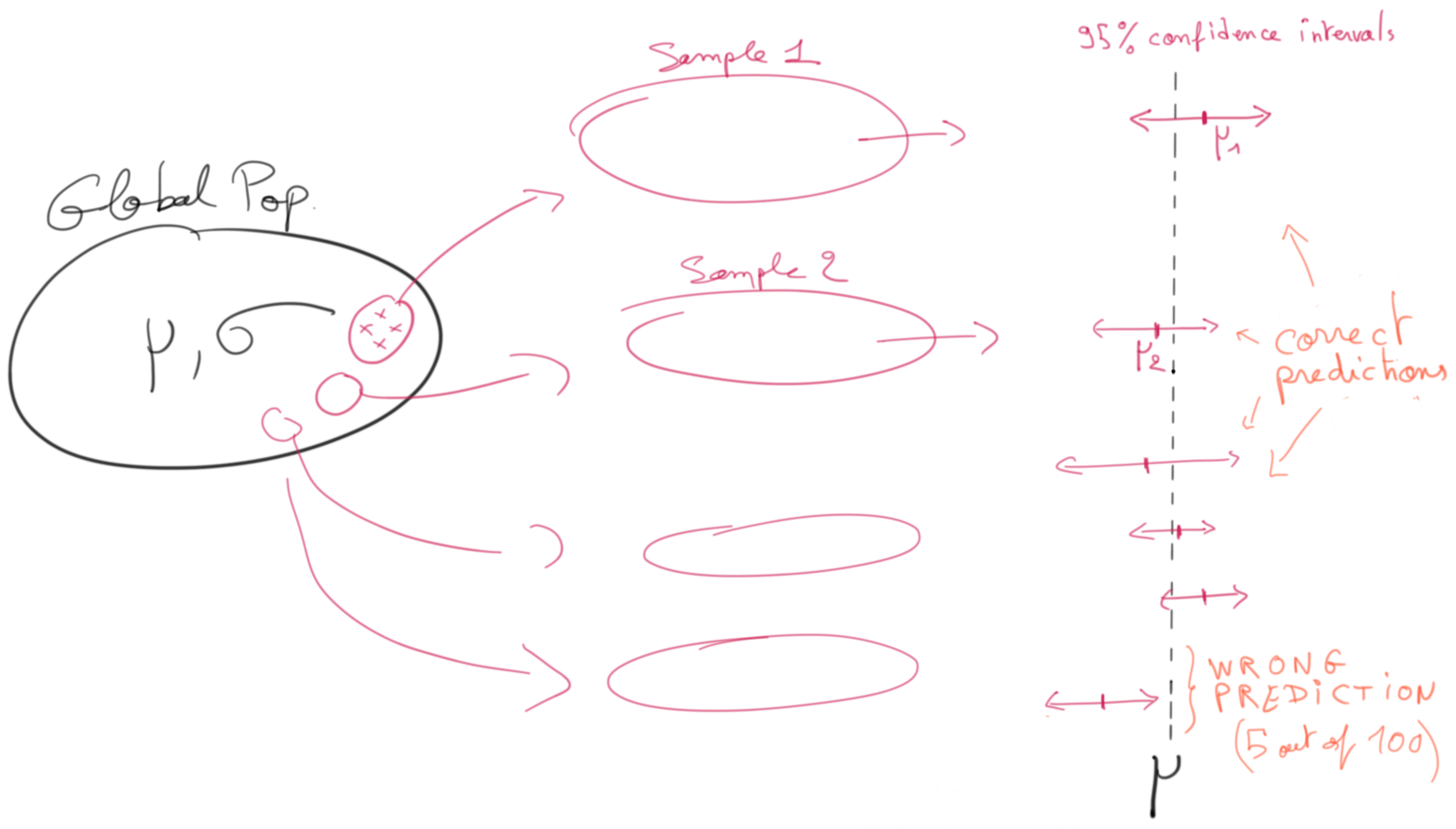
c

m

[95% confidence interval]

### **Confidence Interval (interpretation)**

✅ If we were to repeat this process and construct many other samples, 95% of the intervals produced will actually contain the true US mean pop height



✅ We're 95% confident that [168.7 - 171.2] captures the **true average height**.

❌ Don't say "*there is a 95% probability that*

μ

*is between ...*" because the real

μ

isn't random!

*# We can check these figure using a Cumulative Density Function `cdf`*

**from** **scipy** **import** stats

mu\_estim = stats.norm(170, 0.63)

*# use the cdf to find the probabilities associated with height values*

print('**% c**onfidence interval = ', round(mu\_estim.cdf(171.2) - mu\_estim.cdf(168.7),2))

% confidence interval = 0.95

💡 Actually, there is a formula to find the lower bound and the upper bound of any confidence interval

(ex: 99%)

confidence\_interval = 0.99

sup\_proba = (1 + confidence\_interval)/2 *# 99.5%*

inf\_proba = (1 - confidence\_interval)/2 *# 0.5%*

mu\_upper\_bound = mu\_estim.ppf(sup\_proba)

mu\_lower\_bound = mu\_estim.ppf(inf\_proba)

*# use the inverse of the cdf to find the heights associated with probabilities*

print('mu\_upper\_bound: ', mu\_upper\_bound)

print('mu\_lower\_bound: ', mu\_lower\_bound)

print('**% c**onfidence interval = ', round(mu\_estim.cdf(mu\_upper\_bound) - mu\_estim.cdf(mu\_lower\_bound),2))

mu\_upper\_bound: 171.6227724612358

mu\_lower\_bound: 168.3772275387642

% confidence interval = 0.99

### **Human perception**

Here are some usual English names for confidence intervals   
\_(according to the Intergovernmental Panel for Climate Change: [IPCC](https://archive.ipcc.ch/publications_and_data/ar4/wg1/en/ch1s1-6.html) for instance)\_

* 1-sigma (68%) "likely"
* 90% "very likely"
* 2-sigma (95%) "extremely likely"
* 3-sigma (99.7%) "virtually certain"
* 5-sigma: "proof" threshold in theoretical physics



[Source](https://mirkomazzoleni.github.io/blog/2016/perception_of_probability/)

### **Sample size n considerations**

❓ When is

n

considered **large enough** for the CLT

Three cases:

* If
* n
* >
* 30
* ⇒
* CLT applies **and** the sample std
* s
* can be used to approximate the true pop
* σ
* in z-statistics
* If
* n
* >
* 10
* *and* observations are "non-skewed" and without outliers
* ⇒
* CLT still applies
* If the global population is known to be normally distributed
* ⇒
* CLT always applies even with an arbitrary small
* n
* , in a sense that we can use the Gaussian distribution

❓ When is

n

is **small enough** to consider each draw independent, even without replacement

* n
* <
* 10
* %
* ×
* N

## **4. Hypothesis testing**

### **🥋 Testing a new app feature**

Imagine that I am the PM (Product Manager) for a Social Network Mobile App.

👉 My N = 1000 users spend:

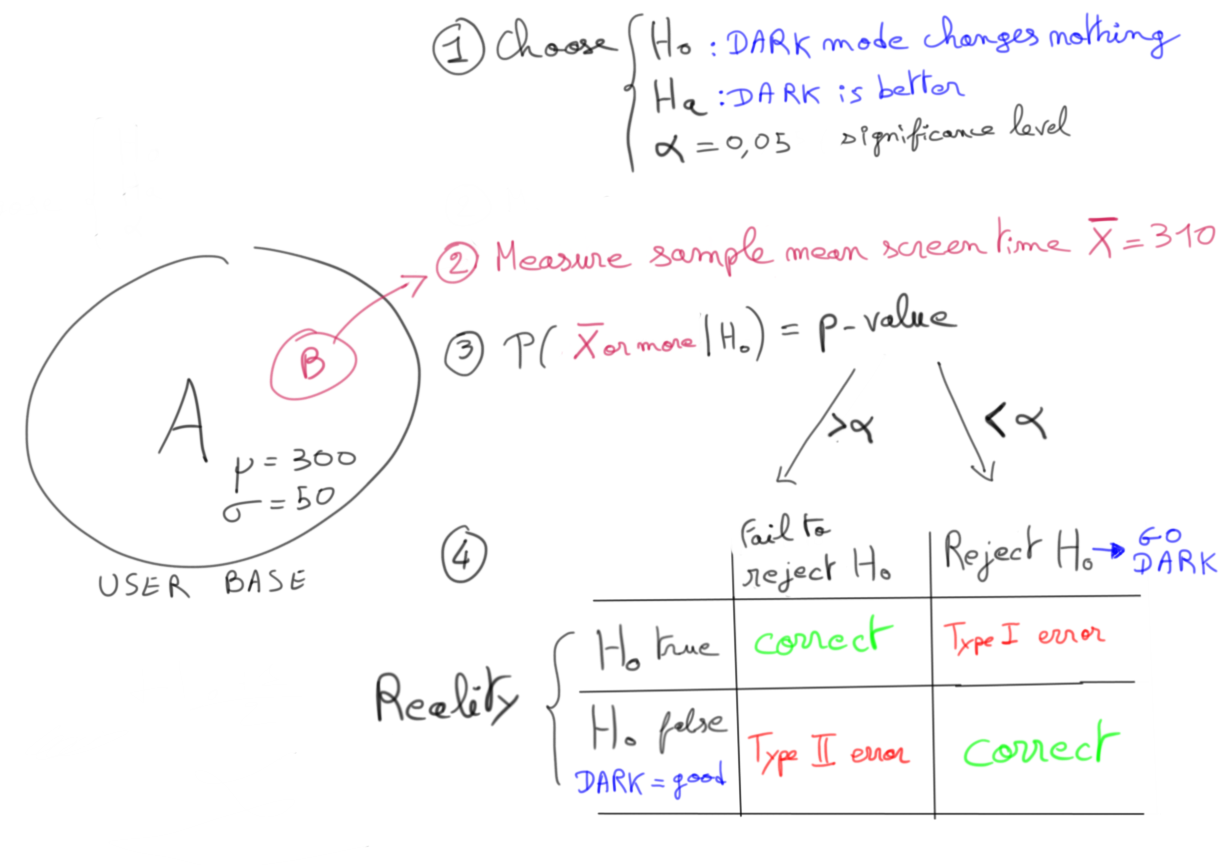
* on average
* μ
* **= 300 seconds** per session
* with a standard deviation of
* σ
* **= 50 seconds**.

💡 I have the intuition that changing the background color from a light to a dark mode would increase the time spent per session.

❓ **How could I test my hypothesis rigorously**, to convince my CTO to roll out the new feature ❓

#### **Step ⓵ : Create a "A/B Test" (the Experiment Design)**

1. Develop the corresponding feature (dark mode)
2. Create two groups (control group vs. treatment group)
3. Randomly assign
4. n
5. **= 100 users** to the treatment group
6. Deploy dark mode only to treatment group
7. Collect behavior statistics: We find a **sample mean**
8. ¯¯¯¯¯¯¯
9. X
10. n
11. **=**
12. ¯¯¯¯¯¯¯¯¯¯¯
13. X
14. 100
15. **= 310 seconds**

**Step ⓶ : Test your hypothesis (= statistical analysis of the outcome)**

#### **Step ⓶ : Test your hypothesis (= Statistical Analysis of the outcome)**

1. Create **Null Hypothesis**
2. H
3. 0
4. :
5. μ
6. *= 300 (unchanged) in dark mode*
7. Create **Alternative Hypothesis**
8. H
9. a
10. :
11. μ
12. *> 300 (increased) in dark mode*
13. Choose a **Significance Level**
14. α
15. for your experiment (ex:
16. α
17. = 5%)
18. Suppose that
19. H
20. 0
21. is true, and compute the probability of observing a sample mean
22. ¯¯¯¯¯¯¯
23. X
24. n
25. ≥
26. 310

👉 This probability is called the **p-value** =

P

(

(

¯¯¯¯¯¯¯

X

n

≥

310

)

|

H

0

)

* If **p-value** <
* α
* , then we **reject** the null hypotheses
* H
* 0
* in favor of the alternative hypothesis
* H
* a
* If **p-value** >
* α
* , then we **fail to reject** the null hypothesis
* H
* 0
* . This doesn't mean we accept
* H
* 0
* ❗️

Let's compute our p-value

Since

n

=

100

is large enough, the CLT applies and tells us that

👉 The distribution of sample means

¯¯¯¯¯¯¯

X

n

should follow the normal distribution:

¯¯¯¯¯¯¯

X

n

≈

N

(

μ

,

σ

√

n

)

Supposing

H

0

is true (unchanged behavior), then we know that

μ

= 300 and

σ

= 50

¯¯¯¯¯¯¯¯¯¯¯

X

100

≈

N

(

300

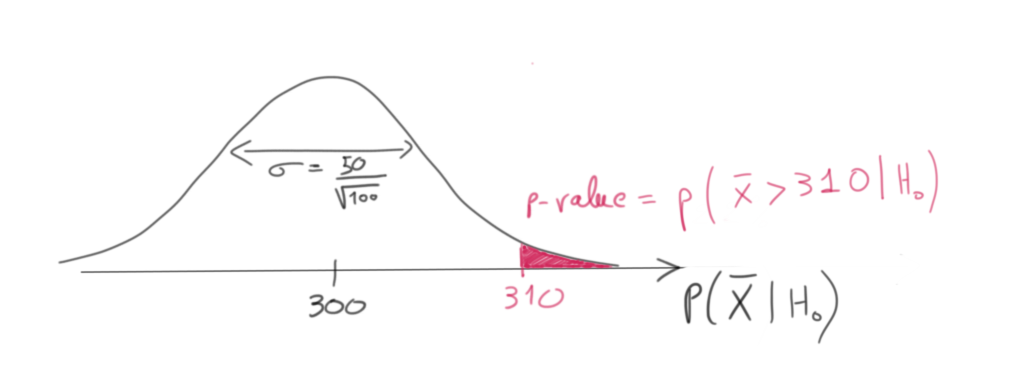
,

50

√

100

)



A standard z-table or a numerical computation would help up compute the p-value

**from** **scipy.stats** **import** norm

X = norm(300, 50/(100\*\*0.5))

p\_value = (1 - X.cdf(310));

round(p\_value,2)

0.02

👍 p-value < 0.05

⇒

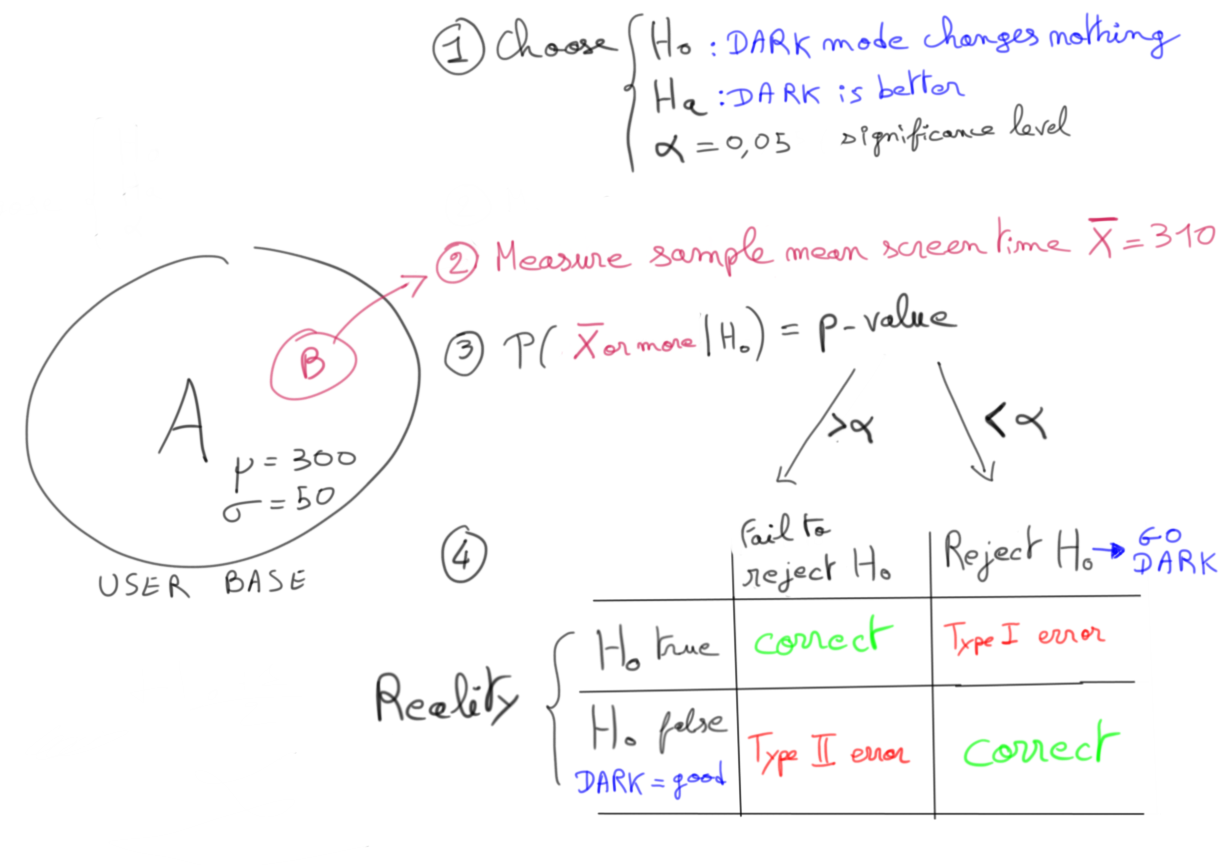
We can safely reject the Null Hypothesis

H

0

⇒

The dark mode is a real plus!



You will also often encounter the **power** of a statistical test (the larger the better):

* Power is the probability that we will **correctly reject the null hypothesis** (if it was correct to reject it)
* Power = Proba of *not missing out* a great feature in A/B testing
* Power = Proba of *not missing out* an effective new drug in clinical trial
* Power = P(not making a type II error)

📺 [StatQuest intuitive video](https://www.youtube.com/watch?v=Rsc5znwR5FA&list=PLblh5JKOoLUIcdlgu78MnlATeyx4cEVeR&index=112&t=0s)

#### **Choosing significance level 𝞪 ?**

α

=

0.05

is the standard significance level we generally start with.

*Notes:*

1. It is the value usually used for clinical trials
2. It can vary from one industry to the other, from one experiment to the other, ...

❌ Never change

α

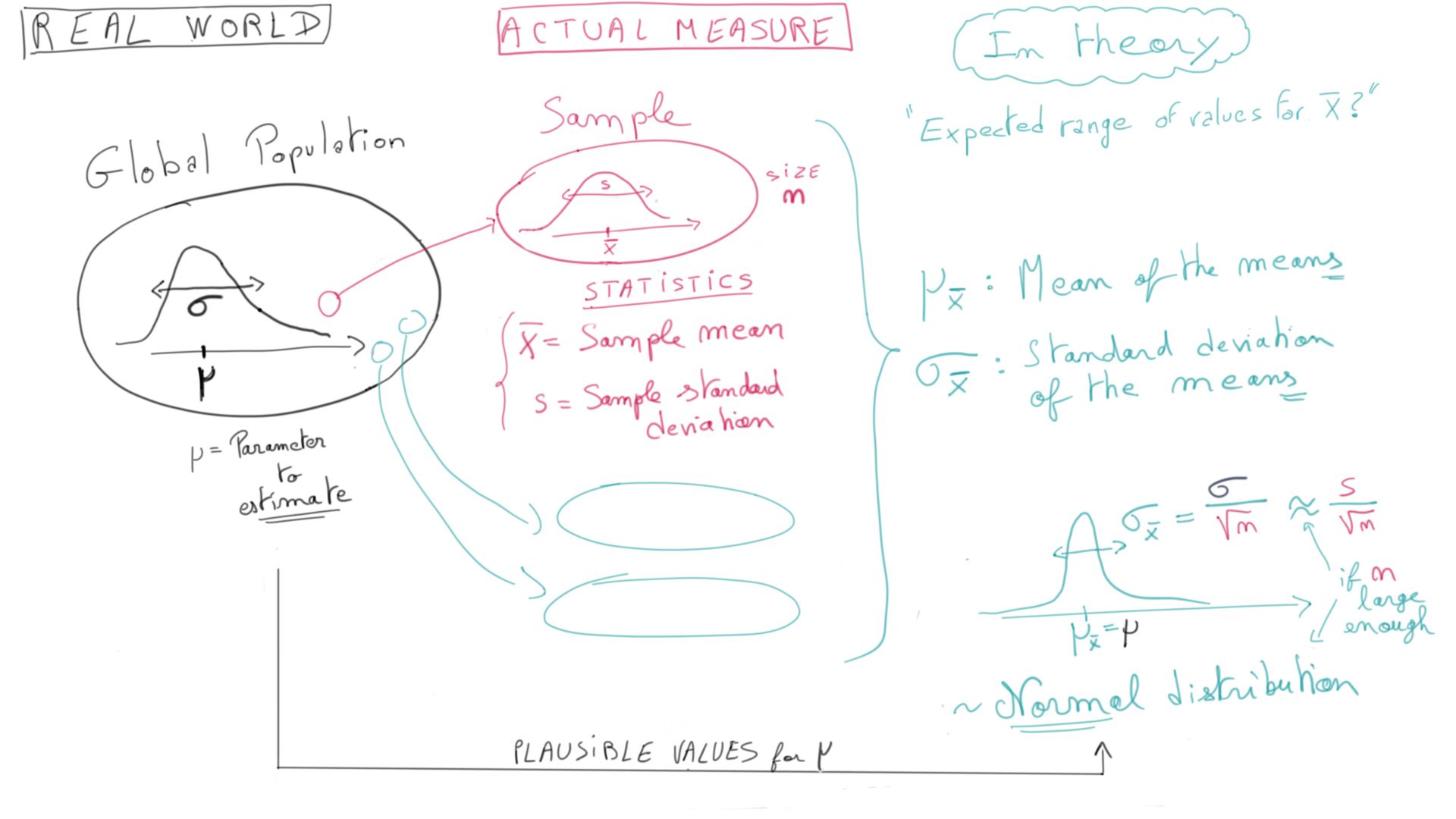
**afterwards** to reject / fail to reject your hypothesis to your own will...

✅ Choose your

α

**beforehands**, depending on your susceptibility to **Type I errors (False Positives)** *vs.* **Type II errors (False Negatives)**

## **5. t-tests (for small sample sizes)**

****

🤔 What to do when the **sample size n is not large enough** and we don't know the true

σ

population ?

Z

=

¯¯¯¯¯

X

−

μ

σ

√

n

* cannot be approximated by
* N
* (
* 0
* ,
* 1
* )
* cannot be computed without knowing true
* σ
* of the population

💡 However we can always compute the **T-statistics**:

T

=

¯¯¯¯¯

X

−

μ

s

√

n

And fortunately we can prove that:

T

=

¯¯¯¯¯

X

−

μ

s

√

n

∼

T

n

−

1

(

0

,

1

)

is called a **Student distribution with n-1 degrees of freedom**

🙌 All good!

✅ Everything applies as before, but replace

N

with

T

:

* Use **t-tests** instead of z-tests
* Compute **confidence interval** using the c.d.f. of a Student distribution
  + Choose the correct number of degrees of freedom!
* **Test Hypothesis** (compute p-value with a significance level
* α
* )

❗️ Still requires *independent* and *random* sampling

### **Student t-distribution**

T

ν

(

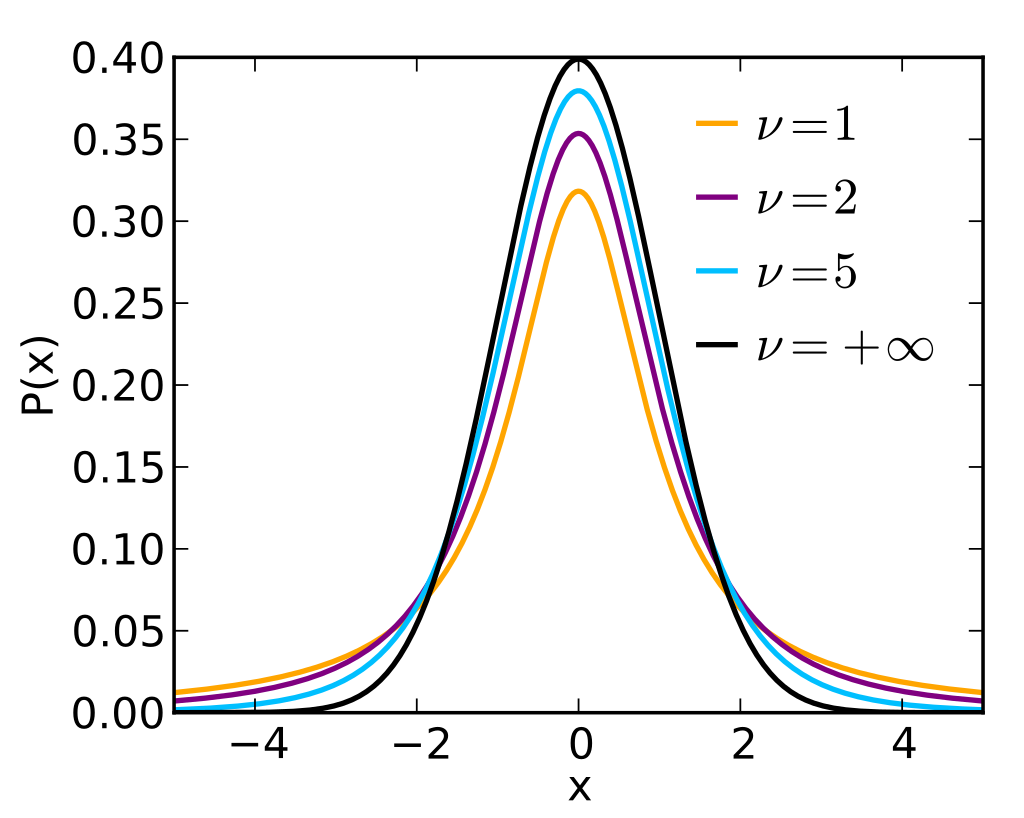
μ

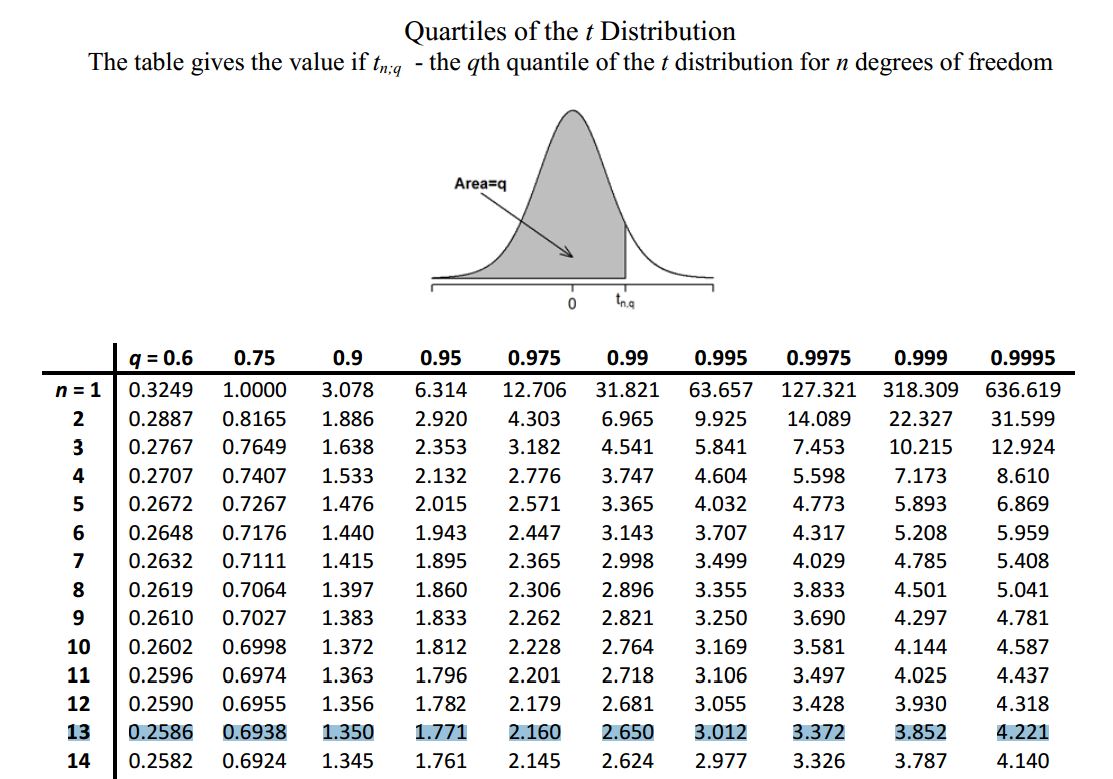
,

σ

)

* One distribution per **degree of freedom**
* ν
  + 📚 [Statistics By Jim - Hypothesis Testing - Degrees of Freedom=](https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/)
* Accessible via scipy.stats.t or via t-table
* "Fatter" tails compared to a
* N
* ormal distribution
* T
* ν
* −−−→
* ν
* →
* ∞
* N





If we were to measure 14 people randomly and to find an average height value with a t-score of 3, then this measured average height would extremely highly (99.5%) improbable!

### **Central Limit Theorem (generalized)**

* X
* 1
* …
* X
* n
* independent random variables sampled from a global pop with mean
* μ
* and std
* σ
* ¯¯¯¯¯
* X
* =
* X
* 1
* +
* ⋯
* +
* X
* n
* n
* the sample mean (also referred to as the *empirical mean*)
* s
* =
* ⎷
* 1
* n
* −
* 1
* n
* ∑
* i
* =
* 1
* (
* x
* i
* −
* ¯¯¯
* x
* )
* 2
* the sample standard deviation

👉 For

n

large enough:

T

∼

Z

=

(

¯¯¯¯¯

X

−

μ

σ

/

√

n

)

∼

N

(

0

,

1

)

👉 For any

n

in

N

:

T

=

(

¯¯¯¯¯

X

−

μ

s

/

√

n

)

∼

T

n

−

1

(

0

,

1

)

🧐 Why (n-1)? Cf. [Bessel's Correction](https://www.statisticshowto.com/bessels-correction/)

## **6. Bayesian Interpretations**

🥋 Let's sample a coin by flipping it

n

times to measure its **fairness** (i.e, the probability

p

(

H

)

=

p

(

μ

,

σ

)

of landing on *Head*).

🤔 Is

μ

equal to 0.5 ? Is the coin fair ?

#### **1. Prior to the experiment, we may have an opinion about the coin fairness**

* This initial belief is called the **prior probability**
* p
* (
* H
* )
* If we have no opinion, we model the
* p
* (
* H
* )
* as a uniform distribution over [0,1]

#### **2. Toss the coin n = 10 times**

* We find sample mean of 0.7 and sample deviation
* s

❓ What is our new best guess for H, after having seen the new data  
❓ i.e What is

p

(

H

|

d

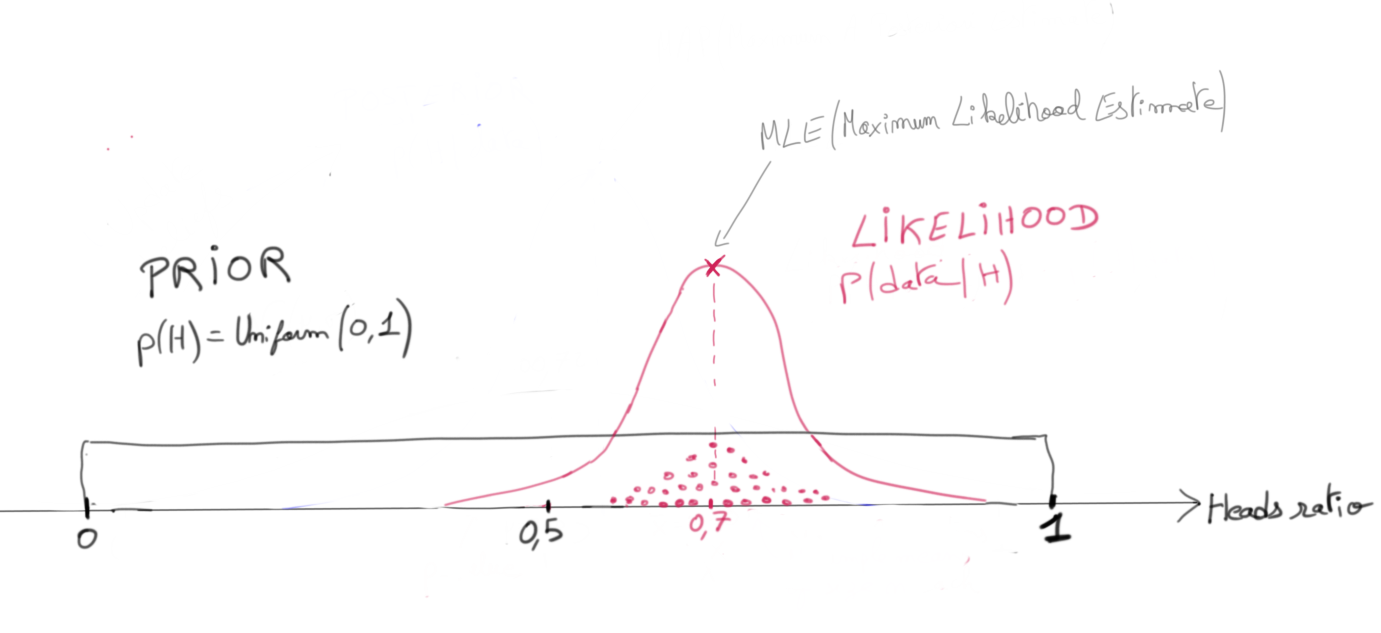
a

t

a

)

**= posterior proba**

****

* In the absence of more **prior beliefs p(H)**, we can rely only on our observation
  + Our most likely estimate for
  + μ
  + is now 0.7 (maximum likelihood estimate MLE)
  + Our new best guess (**posterior proba**) of the coin fairness
  + p
  + (
  + H
  + |
  + d
  + a
  + t
  + a
  + )
  + is equal to
  + N
  + (
  + 0.7
  + ,
  + σ
  + √
  + n
  + )
  + =
  + N
  + (
  + 0.7
  + ,
  + 0.1
  + )
* Indeed, the CLT tells us that the most likely distribution from which such a mean of 0.7 may have been drawn is
* N
* (
* 0.7
* ,
* σ
* √
* n
* )
* . (called the **likelihood** of observing data)

Now, imagine **we do have a prior belief** about the coin's fairness

* Extreme values close to 0 and 1 seem very unlikely to us (we can see it visually)
* Most coins are usually fair, so we think the most probable value is 0.5

👉 We will model our prior belief

p

(

H

)

=

N

(

0.5

,

0.3

)

❓ What is our new posterior proba estimate

p

(

H

|

d

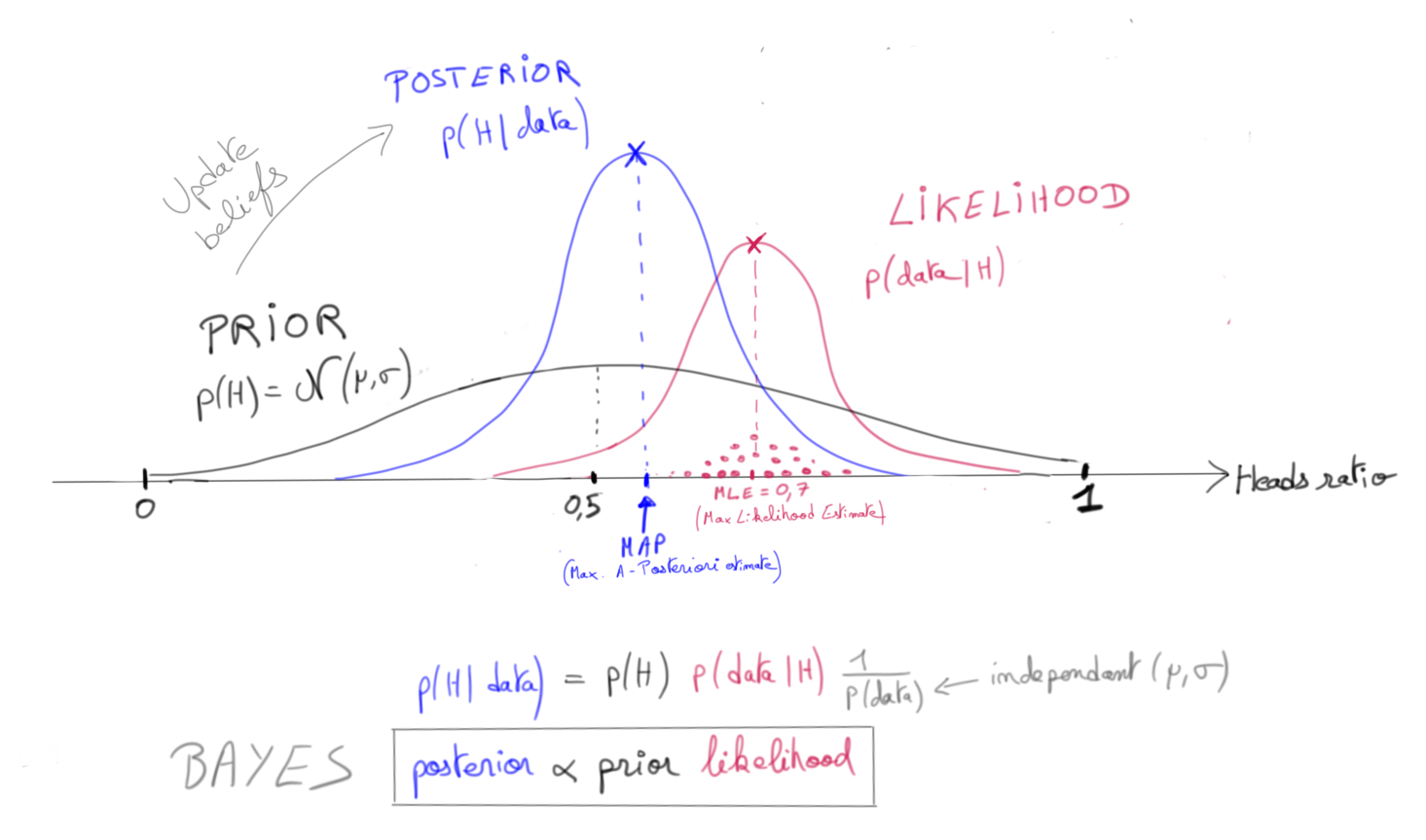
a

t

a

)

= ❓



👉 Use **Bayes** to update our prior belief

p

(

H

)

into our **posterior belief**

p

(

H

|

d

a

t

a

)

## **Bibliography and Videos**

* 📺 [3Blue1Brown - Bayesian Updating and Probability Density Functions](https://www.youtube.com/watch?time_continue=3&v=rhuMH8A5t8s&feature=emb_logo)
* 📚 [Towards Data Science - Jonny Brooks-Bartlett - Bayesian Inference for Parameter Estimation](https://medium.com/towards-data-science/probability-concepts-explained-bayesian-inference-for-parameter-estimation-90e8930e5348)
* 📺 [Khan Academy - Statistics and Probability](https://www.khanacademy.org/math/statistics-probability) (~ 20h)
* 📚 [Miguel A. Hernán and James M. Robins - Causal Inference - What if](https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2019/10/ci_hernanrobins_26oct19.pdf) (300-page M.Sc.level textbook)

## **🚀 Your turn**

🗓 Now:

* Creation of the "Orders" training set
* Quick analysis of the training set with a simple Linear Regression

🗓 Next session:

* In-depth analysis of the "Orders" dataset with a Multivariate Linear Regression